numero 230
ottobre 1987
anno XX
volume XXXIX

LE SCIENZE
edizione italiana di SCIENTIFIC AMERICAN

Stars & Stripes

L'imbarcazione che ha vinto l'edizione 1987 della Coppa America costituisce il risultato di un grande impegno tecnologico e progettuale nel quale i calcolatori elettronici hanno svolto un ruolo fondamentale

di John S. Letcher Jr., John K. Marshall, James C. Oliver III e Nils Salvesen

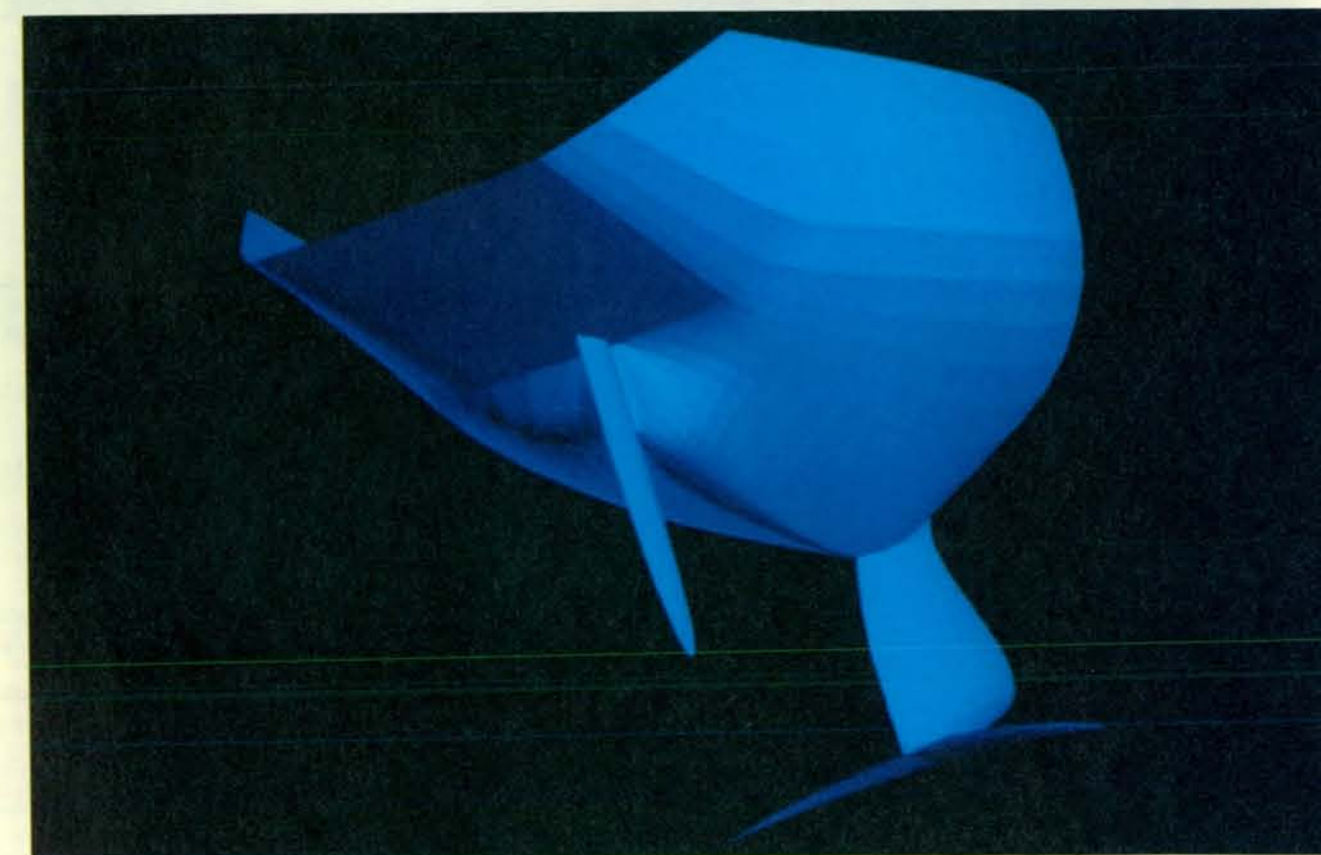
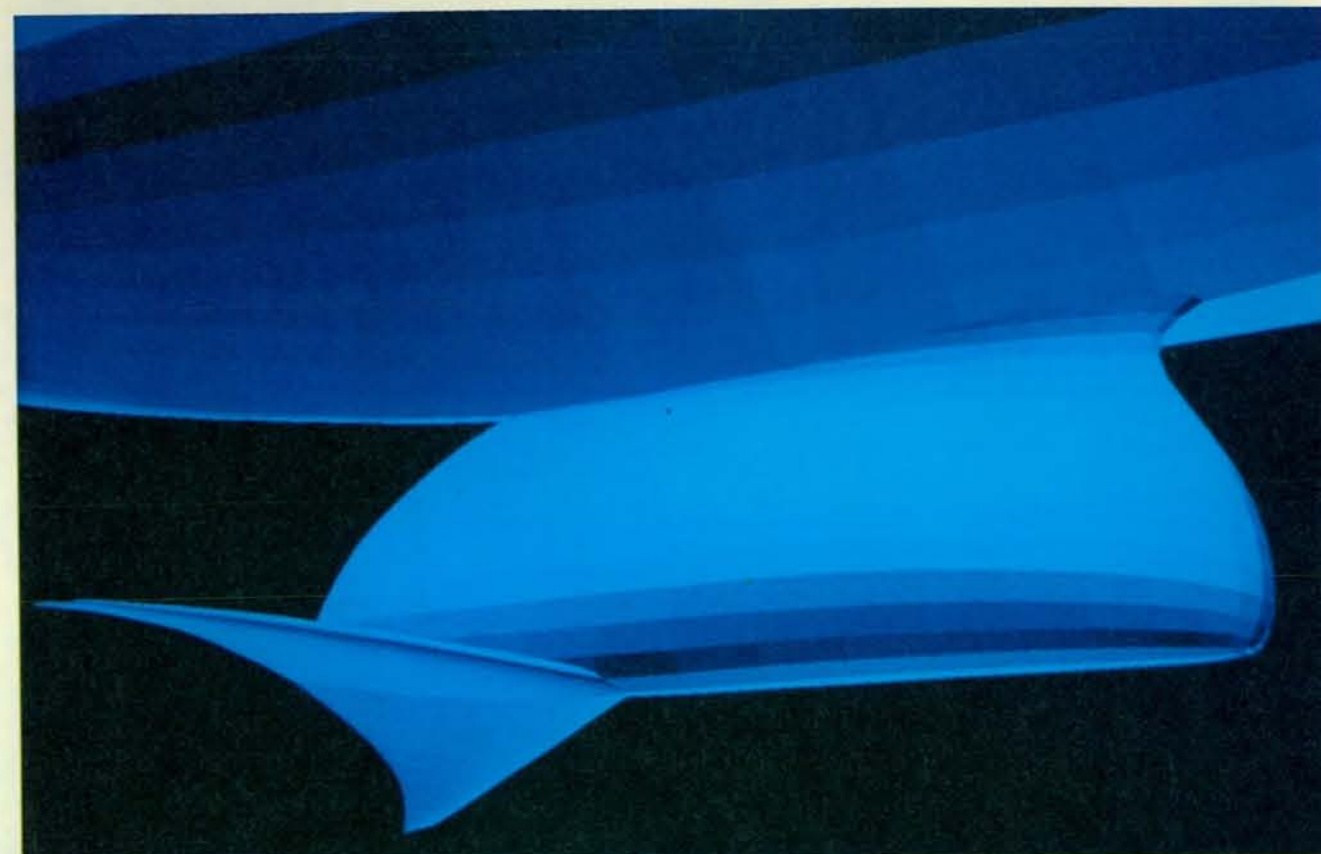
Fino alla vittoria di *Australia II* nella edizione 1983 della Coppa America, negli ambienti nautici si era convinti che non si potesse fare molto di più, dal punto di vista tecnologico, per migliorare le barche della classe 12 metri Stazza Internazionale (SI). [Alle regate per la conquista della Coppa America sono ammesse le imbarcazioni della classe 12 metri Stazza Internazionale (SI) che, d'ora innanzi, chiameremo semplicemente 12 metri. La formula di stazza dei 12 metri fu elaborata nei primi anni di questo secolo ed è un'equazione matematica che, come vedremo più avanti, correla alcune dimensioni dello scafo con il valore della superficie velica. Possiamo dire, quindi, che tutti i 12 metri appartengono a una grande famiglia progettuale nella quale la geometria e i pesi possono variare entro determinate fasce. *n.d.r.*] I risultati ottenuti nelle regate dal 1973 in poi, avevano consolidato l'opinione che quegli scafi fossero ormai vicini all'optimum e che sarebbero stati possibili solo miglioramenti marginali. Eppure *Australia II* nel 1983 vinse soprattutto grazie a radicali innovazioni nel progetto della chiglia.

Ciò spinse Dennis Conner, skipper della barca americana *Liberty* che aveva perso la Coppa, a meditare insieme ai suoi consiglieri la lezione implicita nel successo australiano e a definire una nuova strategia per la preparazione alla sfida del 1987. Essi istituirono un consorzio, la Sail America Foundation, con l'intenzione di riportare la Coppa negli Stati Uniti utilizzando il meglio della tecnologia americana per la progettazione della barca sfidante. Attraverso l'intenso programma di studi teorici e di verifiche portato avanti da Sail America emerse una barca, *Stars & Stripes*, che riportò negli Stati Uniti la Coppa battendo in quattro regate *Kookaburra III*, il difensore australiano del titolo.

Intanto, negli Stati Uniti, erano stati costituiti altri cinque consorzi e sette consorzi, di cinque altri Paesi (Canada, Francia, Inghilterra, Italia e Nuova Zelanda), si unirono alla sfida. Per difendere la coppa entrarono poi in lizza quattro consorzi australiani con sette nuove barche. Ispirandosi ai metodi che avevano condotto al successo *Australia II*, tutti i concorrenti diedero grande importanza allo sviluppo di modelli impiegando elaboratori elettronici per il loro studio idrodinamico. Risultava infatti ben chiaro che la sfida sarebbe stata giocata in pari misura fra i tecnologi e gli equipaggi dei vari paesi. Nel 1987 nessuna barca avrebbe potuto vincere senza essere all'avanguardia dal punto di vista tecnologico, ma anche una barca così avanzata avrebbe sicuramente perduto senza un equipaggio di prim'ordine.

Già all'inizio del 1984, Sail America riunì una squadra di progettisti, guidata da uno di noi (Marshall), con l'incarico di integrare il tradizionale lavoro del progettista di imbarcazioni con i più recenti sviluppi della tecnologia. Facevano parte del gruppo tre dei maggiori esperti statunitensi del settore: Britton Chance, Jr., Bruce Nelson e David Pedrick. Lo studio tecnologico ha richiesto un lavoro equivalente a circa 10 anni-uomo da parte di una trentina tra ricercatori e ingegneri. Gran parte di esso è stato svolto dalla Science Applications International Corporation (SAIC), sotto la direzione di uno di noi (Salvesen), e da un gruppo di ingegneri della Grumman Corporation, diretto da Charles W. Boppe.

Il gruppo di tecnici si è misurato con un compito di progettazione caratterizzato da due condizioni tipiche delle regate dei 12 metri. La prima condizione è che le barche in competizione non sono perfettamente identiche, contrariamente a quanto avviene in molte altre regate, fra le quali quelle dei Giochi Olimpici . Da ciò deriva che i progettisti dei 12 metri possono sfruttare al meglio le loro risorse tecnologiche. La seconda condizione è costituita dal fatto che i 12 metri, a differenza dei monotipi, regatano sfidandosi a coppie. Questo modo di regatare, che dà come risultato un vincente e un perdente, si chiama *Match Race*. Inoltre il percorso della regata privilegia la navigazione di bolina; si deve infatti risalire il vento in quattro degli otto tratti del percorso, pari a più di metà della sua lunghezza totale di 24 miglia nautiche. Una barca a vela non può navigare direttamente controvento, ma deve procedere eseguendo una serie di bordi di bolina, durante la quale la barca vira, cambiando direzione e ricevendo il vento prima da un lato e poi dall'altro. La regata a coppie e di bolina premia l'abilità e le capacità tattiche dell'equipaggio, oltre alla velocità pura della barca.

La progettazione di un 12 metri è soggetta all'omonima regola di stazza, formulata e controllata dalla International

Queste immagini al calcolatore mostrano la configurazione della chiglia di *Stars & Stripes* (*in alto*) e dell'intera opera viva (*in basso*). Oltre alla struttura ad alette all'estremità inferiore, la chiglia possiede un piccolo alettone (trimmer, non visibile nell'illustrazione) che viene controllato dal pozzetto con una ruota distinta da quella del timone. Per scegliere la chiglia di *Stars & Stripes*, la Sail America Foundation ha valutato al calcolatore centinaia di configurazioni diverse. Queste immagini sono state eseguite da Design Systems & Services, Inc., Science Applications International Corporation e Hewlett-Packard Company. Esse. mostrano la configurazione finale della chiglia di *Stars & Stripes*

Yacht Racing Union di Londra. La regola è rappresentata da una formula matematica molto semplice, la cui interpretazione è però abbastanza complicata tanto da riempire un fascicolo di 25 pagine scritte fittamente. La regola consiste di una relazione matematica fra alcune dimensioni dello scafo e delle vele, fra le quali si possono fare le scelte ritenute più opportune. Per esempio, si può optare per uno scafo più lungo, ma allora occorre compensare il vantaggio riducendo la superficie delle vele. La regola pone anche dei limiti a certe dimensioni dell'imbarcazione, come un valore minimo del baglio massimo (la larghezza dello scafo), un pescaggio massimo (profondità di immersione), un peso minimo, che aumenta con l'aumentare della lunghezza al galleggiamento dello scafo, e un'altezza massima per le varie vele.

La regola consente la realizzazione di una certa gamma di 12 metri. A un'estremo della gamma si trovano scafi corti e leggeri, caratterizzati da grande superficie velica. Tali imbarcazioni sono adatte a regatare con venti leggeri. All'altro estremo vi sono scafi lunghi, con chiglie più pesanti e meno invelati, che si comportano bene in presenza di venti forti.

Dato che la sfida del 1987 si doveva tenere per la prima volta nelle acque dell'oceano Indiano, al largo di Perth, in Australia, i progettisti hanno dovuto affrontare i problemi conseguenti alle particolari condizioni atmosferiche locali. Soltanto quattro fra gli sfidanti sarebbero emersi dalle regate di qualificazione, previste per i mesi di ottobre, novembre e dicembre 1986, durante la primavera australe caratterizzata da venti in genere più leggeri rispetto all'estate australe. Occorreva quindi progettare una barca in grado di superare le eliminatorie con i venti primaverili moderati e continuare a eccellere con i forti venti estivi del gennaio 1987. I regolamenti richiedono che venga impiegata la medesima barca per l'intero ciclo delle regate, ma consentono di cambiare la chiglia e le vele, a patto che non si esca dalla stazza dei 12 metri.

Regatare con i forti venti estivi al largo di Perth pone notevoli problemi sia all'equipaggio, sia ai progettisti della barca. Venti con velocità media da 20 a 25 nodi (da 40 a 50 chilometri all'ora) impongono agli uomini e alle attrezzature uno sforzo più che doppio rispetto alle brezze di 12 nodi tipiche di Newport, sede delle nove precedenti edizioni della Coppa America svoltesi dopo la seconda guerra mondiale. Poco si sapeva sul come «portare» i 12 metri in tali condizioni. Una chiglia con alette del tipo che era risultato utile per *Australia II* avrebbe ancora presentato qualche vantaggio? O sarebbe stato meglio ideare una configurazione completamente diversa per il complesso chiglia-timone? Pareva ovvio che per navigare con venti forti fosse meglio sacrificare la superficie velica in favore di uno scafo più lungo e pesante, ma quanto ci si doveva spingere in tale direzione? Il mare agitato al largo di Perth avrebbe consigliato la scelta di nuove forme dello scafo per ridurre la maggiore resistenza all'avanzamento? Sarebbe valsa la pena di costruire una barca più stabile aumentando il baglio massimo o il dislocamento? Nessuno conosceva le risposte a queste domande e non v'era molto tempo per trovarle.

I nostri sforzi per rispondere in maniera adeguata ci hanno portato ad affrontare svariate discipline e molti campi della tecnologia marittima. Ogni indagine al calcolatore aumentava la nostra capacità analitica e decisionale. I dati meteorologici e oceanografici costituivano la colonna portante del nostro lavoro: dovevamo conoscere dettagliatamente le condizioni che si sarebbero incontrate sul campo di regata. Le prove in vasca fornivano altri dati fondamentali. Sono state condotte 40 prove con modelli in scala 1 : 3 e sono stati sviluppati nuovi metodi per l'elaborazione dei dati ottenuti. Sono stati anche raccolti e analizzati dati sulle prestazioni reali in acqua di cinque imbarcazioni da 12 metri.

Per ottimizzare la forma dello scafo abbiamo tenuto conto delle teorie note sulla resistenza offerta dalle onde al moto di uno scafo. Sono stati impiegati programmi per la determinazione del flusso dei fluidi intorno a corpi di forma complessa per analizzare il comportamento idrodinamico dello scafo e della chiglia (la cosiddetta opera viva) al fine di valutare e migliorare centinaia di configurazioni. I metodi usuali per determinare le prestazioni di uno scafo in acqua calma sono stati estesi per prevederne il comportamento sotto vela in presenza di onde e nel corso delle manovre. Per confrontare il comportamento di centinaia di possibili sfidanti contro i loro ipotetici contendenti sono state impiegate la teoria della probabilità, la teoria dei giochi e simulazioni capaci di rappresentare le variazioni di posizione e di velocità di una barca in regata.

Il lavoro di progettazione era incentrato su un programma di simulazione al calcolatore della dinamica di un'imbarcazione a vela denominato Velocity Prediction Program (VPP). Programmi di questo tipo sono stati sviluppati negli anni settanta da diversi ricercatori. Per il nostro progetto, uno di noi (Oliver) ha sviluppato un VPP appositamente concepito per la progettazione di un 12 metri e la simulazione delle regate. Fornendo informazioni sulla struttura fisica della barca si ottengono previsioni degli angoli di rotta e delle velocità per qualsiasi combinazione della velocità reale del vento (la sua velocità rispetto all'acqua) e della rotta della barca rispetto alla direzione del vento.

Fra i fattori considerati dal VPP vi sono il vento reale e il vento apparente, che è quello che incide sulle vele. Esso considera pure le forze aerodinamiche prodotte dalle vele e, in minor misura, dalla parte emersa dello scafo (l'opera morta). Queste forze sono la resistenza, parallela al vento apparente, e la portanza, perpendicolare a questo (*si veda l'illustrazione nella pagina successiva*). Il VPP tiene anche conto delle forze idrodinamiche agenti sull'opera viva: ancora una resistenza, diretta questa volta in senso contrario al moto della barca, e la spinta laterale, perpendicolare alla prima. Perché la barca abbia una velocità costante queste quattro forze devono avere risultante nulla. Inoltre anche la somma dei momenti che agiscono sulla barca, tendendo a farla ruotare, deve essere zero. In particolare i momenti stabilizzanti (dovuti alla zavorra e alla forma dello scafo) devono bilanciare quelli di rovesciamento (dovuti alle forze esercitate dal vento sulle vele), producendo uno sbandamento costante e non eccessivo.

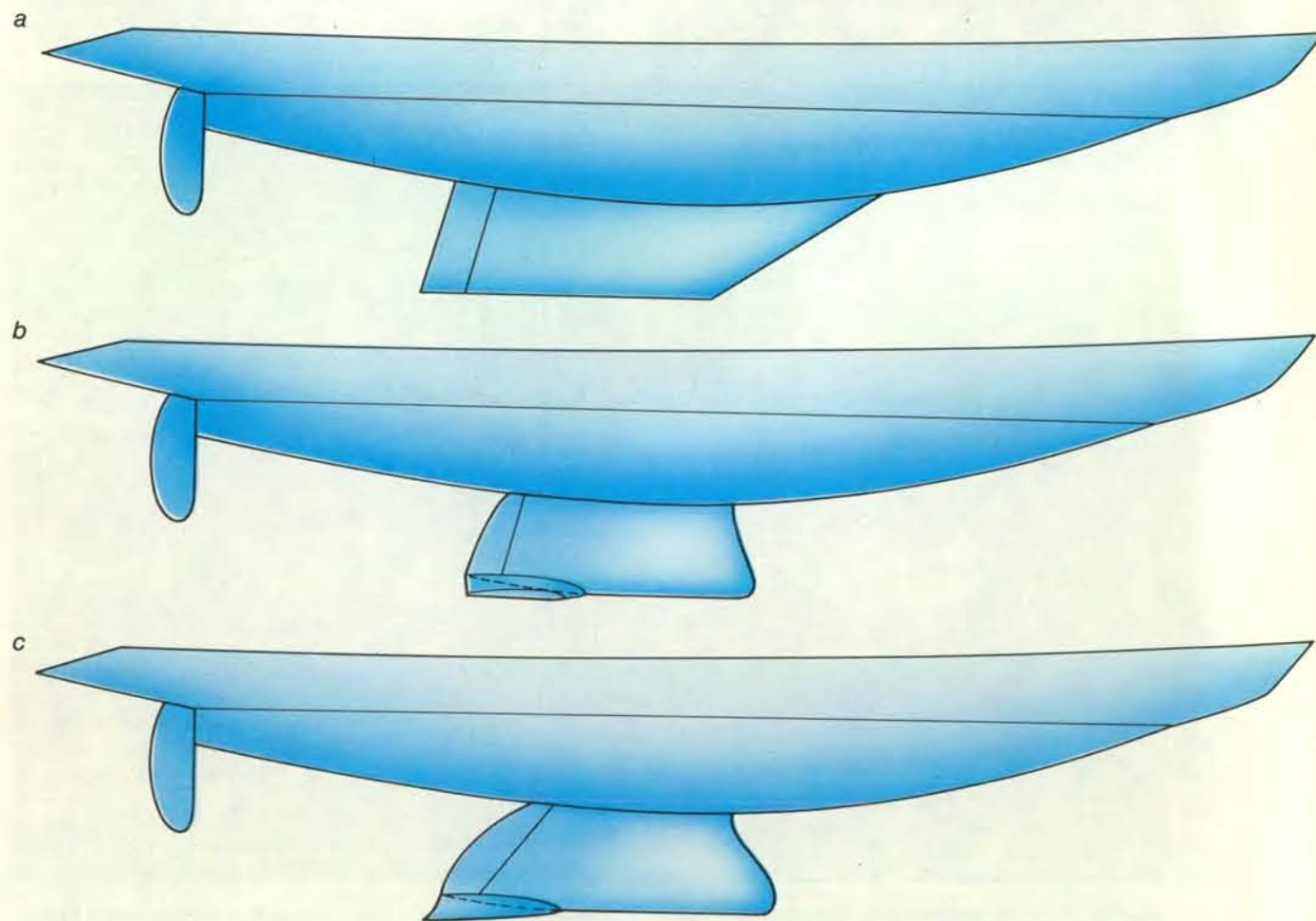
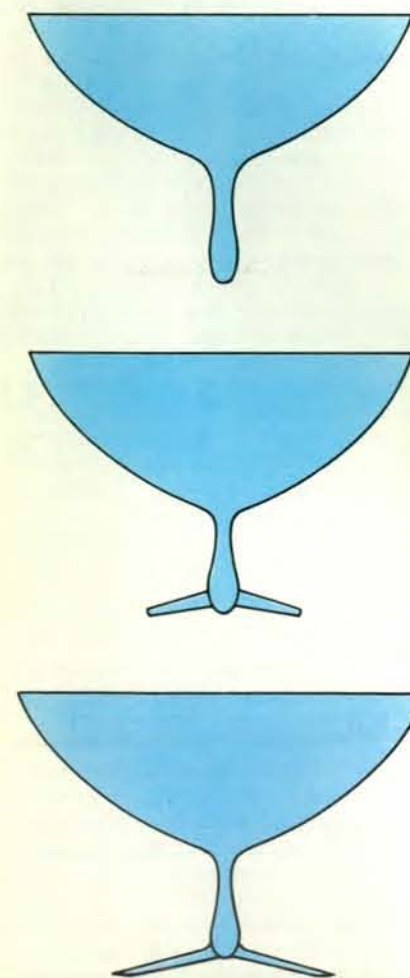
Queste forze e questi momenti dipendono tutti dalla velocità e dalla direzione del vento reale, dall'angolo di sbandamento e dalla velocità della barca. Dipendono anche da diverse variabili che definiscono la disposizione e la regolazione delle vele e che il programma ottimizza quando ricerca l'equilibrio delle forze e dei momenti.

Il programma può contemporaneamente cercare la direzione di navigazione (rotta) più favorevole. Uno skipper che deve risalire il vento cerca sempre il miglior compromesso fra una rotta che consenta di stringere il più possibile il vento e una che mantenga elevata la velocità della sua barca. «Poggiando», ovvero allargando la rotta rispetto al vento reale, la barca guadagna in velocità perché aumenta la componente della portanza nella direzione del moto. Questa manovra ha però un costo: risulta necessario coprire un percorso più lungo per raggiungere un determinato punto, risalendo il vento.

Al contrario, «orzando», cioè stringendo il vento, si diminuisce la lunghezza della rotta da percorrere per raggiungere il punto voluto, ma si rallenta la velocità della barca. Esiste quindi una rotta ottimale, seguendo la quale è massima la componente della velocità nella direzione del vento. Prendendo questa rotta la barca raggiungerà nel tempo minimo un determinato punto sopravento. Una situazione simile si verifica navigando con il vento in poppa. Per ogni velocità del vento vi è una certa direzione rispetto al vento reale - generalmente meno di 180 gradi, cioè con il vento non perfettamente in poppa, ma un po' di lato (al giardinetto) - che rende massima la componente della velocità nella direzione del vento. Il programma VPP aiuta a trovare queste rotte ottimali.

Il modello matematico della navigazione a vela sul quale si basa il VPP è ovviamente importante per la progettazione delle imbarcazioni a vela fino a che è in grado di dare risposte corrette alle domande del progettista. Un'elaborazione ha un costo irrisorio (meno di 15 dollari) e impegna un tempo limitatissimo se confrontato con una prova idrodinamica in vasca (25 000 dollari), per non parlare di una prova con un'imbarcazione vera (da 500 000 a un milione di dollari). Il calcolatore può analizzare in poco tempo un gran numero di idee e di modifiche. La precisione della soluzione matematica consente poi di evidenziare anche minime differenze fra le prestazioni, consentendo di procedere sistematicamente all'ottimizzazione dei progetti. Il VPP fornisce risposte quantitative a una miriade di richieste per le quali precedentemente il progettista poteva far ricorso solamente alla propria intuizione e all'esperienza maturata.

Uno degli elementi più importanti del progetto è costituito dalla chiglia: *Australia II* aveva vinto proprio grazie a un nuovo tipo di chiglia. Naturalmente il nostro obiettivo era innanzitutto di eguagliare i risultati degli australiani e poi di migliorarli, studiando nel corso del lavoro altre configurazioni fuori del comune che potessero offrire qualche vantaggio. A questo scopo abbiamo usato sia il calcolo, sia la sperimentazione. Alla fine crediamo di avere compreso in modo soddisfacente il comportamento della chiglia con alette ideata per *Australia II* dal progettista australiano Ben Lexcen. Malgrado abbiamo pensato e provato altri tipi di appendici, nessuno di questi ci è parso un possibile miglioramento dell'idea di Lexcen.

La chiglia di una barca a vela ha due funzioni complementari: serve come zavorra e come ala immersa che produce la spinta idrodinamica necessaria per bilanciare le forze laterali dovute alle vele. Come zavorra, la chiglia stabilizza la

La configurazione dello scafo e della chiglia di un 12 metri Stazza Internazionale è limitata dalla regola di stazza della classe 12 metri, che consente una notevole libertà nella progettazione della chiglia. Fino al 1983 la chiglia standard aveva una forma trapezoidale rastremata (*a*). Quest'anno *Australia II* (*b*) ha vinto la Coppa con una nuova chiglia con rastrematura inversa e presenza di alette. *Stars & Stripes* (*c*) presenta una rastrematura inversa più accentuata e alette di apertura maggiore. I dettagli della chiglia non sono ancora stati rivelati.

barca nei confronti del momento di sbandamento dovuto alla forza del vento sulle vele. Nei 12 metri la zavorra è di piombo e costituisce il 70 o l'80 per cento del peso totale dello scafo. La funzione fondamentale della zavorra è quella di determinare la posizione verticale del baricentro dell'imbarcazione, che risulta tanto più stabile quanto più il baricentro è basso.

Per raggiungere tale obiettivo, è possibile cambiare la tradizionale forma trapezoidale della chiglia (con la base maggiore in alto) costruendola di forma rettangolare o anche con rastrematura inversa (con la base maggiore in basso). In effetti, la rastrematura inversa è stata un'altra innovazione apportata alla chiglia di *Australia II*. Il baricentro della barca può venire abbassato anche ispessendo il profilo della chiglia nella sua parte più immersa. Può anche risultare utile dare all'estremità inferiore della chiglia la forma di un bulbo affusolato.

Dal punto di vista idrodinamico la chiglia ha poi effetti importanti su quattro dei cinque tipi di resistenza incontrati da uno scafo in movimento. Innanzitutto, se non è ben disegnata o raccordata con lo scafo, produce un'eccessiva resistenza di forma (dovuta a pressioni che si oppongono al moto a causa del distacco dei filetti fluidi). In secondo luogo, dato che ha una notevole superficie bagnata ed è poco estesa nella direzione del flusso, la chiglia contribuisce in maniera notevole alla resistenza d'attrito. In terzo luogo, il volume della chiglia che si muove a poca profondità aumenta la resistenza dovuta alla formazione di onde da parte della barca. Infine, la forma della chiglia ha un effetto importante sull'entità della resistenza indotta, dovuta alla spinta idrodinamica laterale. (Il quinto tipo di resistenza è dovuto alle onde di superficie incontrate dalla barca.)

La resistenza indotta, che costituisce il maggior contributo della chiglia alla resistenza totale all'avanzamento della barca, può venire considerata come il prezzo da pagare per produrre una portanza utile. Un'ala che produce una portanza costante trasferisce continuamente quantità di moto al fluido che la circonda. La quantità di moto si conserva nella scia dell'ala e la perturbazione del fluido si propaga sottovento per un lungo tratto. L'energia cinetica della perturbazione deve venire pagata dal lavoro fatto dall'ala. E la forza che viene vinta con tale lavoro è appunto la resistenza indotta.

La resistenza indotta aumenta con il quadrato della portanza (se la portanza raddoppia, la resistenza quadruplica) e diminuisce con l'inverso del quadrato della lunghezza dell'ala (se la lunghezza raddoppia, la resistenza diminuisce di quattro volte). Si può quindi diminuire la resistenza indotta aumentando la profondità della chiglia, ma la regola di stazza dei 12 metri non lascia molte possibilità in questa direzione. Quindi si sono dovuti prendere in considerazione sistemi non planari. In aerodinamica si sapeva da tempo che superfici portanti multiple possono, in determinate circostanze produrre una resistenza inferiore a quella di una superficie singola della medesima apertura. Un esempio è costituito dal biplano; un altro esempio, che presenta molte somiglianze con la chiglia con alette, è costituito dal monoplano con deflettori (piccole superfici perpendicolari all'ala disposte alle sue estremità). È stato proprio studiando le superfici portanti multiple che abbiamo definito il progetto finale della chiglia con alette di *Star & Stripes* (un progetto che non possiamo ancora rendere pubblico).

Boppe e collaboratori hanno preso l'iniziativa di applicare al discusso problema della progettazione della chiglia i metodi utilizzati nell'industria aerospaziale per prevedere il comportamento aerodinamico di superfici complesse. Essi hanno calcolato la portanza e la resistenza indotta di diverse configurazioni di chiglia con alette, identificando quelle più promettenti, che sono state successivamente verificate con l'impiego di modelli e del VPP.

La responsabilità principale dello studio della resistenza prodotta dalle onde è stata assunta da un gruppo di ricercatori della SAIC guidato da Carl A. Scragg, che si era specializzato in questo settore. Essi hanno cominciato con lo studiare tre scafi di 12 metri, che erano stati sottoposti a prove in vasca, usando un programma basato sulla cosiddetta teoria del bastimento sottile sviluppata da Francis Noblesse del David Taylor Naval Ship Research and Development Center, presso Washington. I primi risultati furono scoraggianti perché le resistenze, calcolate in funzione della velocità, erano molto maggiori dei valori osservati in vasca e il caratteristico andamento in salita della curva della resistenza iniziava a velocità molto più basse.

Scragg ha comunque osservato che, nonostante i grossi errori fatti nei calcoli, le differenze nella resistenza mostrate dai tre scafi venivano previste con notevole accuratezza. Ulteriori prove confermarono i primi risultati. Scragg quindi concluse che, se il programma prevedeva che uno scafo si comportasse meglio di un altro in un dato campo di velocità, questo comportamento doveva venire confermato dalle prove dei modelli.

Pertanto, la teoria del bastimento sottile, malgrado risultasse poco accurata nel calcolo dei valori assoluti della resistenza, è divenuta uno dei fondamenti della progettazione dello scafo. Con il suo aiuto sono stati sviluppati scafi con forme che presentavano una resistenza ridotta alle più importanti velocità di navigazione, cosa che ha contribuito non poco al successo di *Stars & Stripes*.

Si ritiene anche che la barca vincitrice abbia beneficiato dell'attenzione che è stata dedicata alla resistenza dovuta alla viscosità dell'acqua (attrito superficiale) e che tipicamente costituisce il 37 per cento della resistenza complessiva che incontra uno scafo in movimento. Sono disponibili molti rivestimenti che vantano la capacità di ridurre la resistenza viscosa. Thomas G. Lang della Semi-Submerged Ship Corporation ha analizzato più di 30 di tali prodotti, ma non ha trovato nulla che fosse migliore di una superficie dipinta, pulita e liscia. In seguito la 3M ha fornito un rivestimento plastico sperimentale, progettato per l'industria aeronautica, con la superficie solcata da microscopici solchi paralleli. Sulla base dei rapporti pubblicati dalla National Aeronautics and Space Administration sono state scelte profondità e distanze dei solchi ottimali per le condizioni del flusso d'acqua intorno allo scafo di *Stars & Stripes* e sono state fabbricate lastre di questo materiale appositamente concepite per la barca. Malgrado vari problemi riguardanti l'allineamento dei solchi con le diverse direzioni di scorrimento dei filetti fluidi e la formazione di fenditure fra le lastre adiacenti, pensiamo che questa tecnologia abbia consentito una riduzione dal 2 al 4 per cento della resistenza viscosa.

S = SPINTA LATERALE (PORTANZA)
R = RESISTENZA IDRODINAMICA
FORZE IDRODINAMICHE
L = PORTANZA
D = RESISTENZA AERODINAMICA
FORZE AERODINAMICHE
VELOCITÀ DEL VENTO APPARENTE
VELOCITÀ DEL VENTO REALE
ANGOLO DEL VENTO REALE
VELOCITÀ DELLA BARCA
V_U (VELOCITÀ UTILE)

Il comportamento di una barca a vela è determinato dall'equilibrio delle forze idrodinamiche e aerodinamiche. Le prime, generate dalla parte sommersa dello scafo, sono costituite dalla resistenza idrodinamica (*R*), che agisce lungo la direzione di moto, ma in verso contrario, e dalla spinta laterale (*S*), perpendicolare alla prima. Le seconde, generate dalle vele, comprendono la resistenza aerodinamica (*D*), parallela al vento apparente, e la portanza (*L*), perpendicolare a esso. Il vento apparente è la somma vettoriale del vento reale e del vento relativo, dovuto al moto della barca. Il Velocity Prediction Program di Sail America ottimizza l'equilibrio di queste forze.

Il progetto dell'intero complesso delle vele di *Stars & Stripes* è stato semplificato e realizzato in modo più rapido avvalendosi dell'ausilio del calcolatore. In passato i progetti venivano realizzati manualmente e la versione finale, in grandezza naturale, doveva essere preparata in un ampio capannone privo di pilastri. Il procedimento di progettazione al calcolatore è conosciuto con il nome di tracciatura al calcolatore.

Un altro problema affrontato è stato quello meteorologico. In conseguenza degli importanti compromessi necessari per regatare ugualmente bene con venti forti e venti leggeri, era di importanza vitale disporre di previsioni accurate delle condizioni meteorologiche per i mesi fra ottobre e febbraio. La barca vincente doveva venire accuratamente ottimizzata per le condizioni atmosferiche dominanti. Inoltre anche la scelta delle vele e perfino la tattica per la regata del giorno dopo sarebbero state dettate dalle previsioni meteorologiche giornaliere.

R. Leland Davis e Chris Bedford dei Galson Technical Services hanno raccolto e analizzato i dati storici sulle condizioni del vento e del mare. Nel corso della progettazione avevamo come riferimento solo le registrazioni storiche del tempo, fondamentali per diverse scelte progettuali, in particolare per decidere la lunghezza da dare alla barca. Durante le regate le previsioni dovevano venire fatte su diverse scale di tempo: settimane per decidere un cambio di chiglia fra una serie di regate e l'altra, un giorno per decidere se chiedere una sospensione, un paio d'ore per scegliere le vele della giornata dal corredo di circa 150 vele e minuti per prevedere il prossimo salto di vento durante una regata. Il nostro servizio di previsioni utilizzava sia modelli statistici e dinamici, sia dati provenienti da radar meteorologici a effetto Doppler dislocati lungo la costa.

Un altro lavoro importante è stato costituito dalle prove in mare con barche da regata. Sotto la direzione di Robert Hopkins, Jr. di Sail America, è stato portato a termine un ambizioso programma di rilevamenti, prima alle Hawaii, per dieci mesi a partire dall'ottobre 1985, e poi in Australia, per cinque mesi, durante il periodo delle regate. La strumentazione e i programmi di calcolatore per questo lavoro sono stati forniti da Richard C. McCurdy della Ockam Instruments, Inc. I trasduttori posti sulle

barche misuravano la velocità, il vento apparente e quello reale, l'angolazione del timone e del trimmer (un piccolo deflettore sul bordo d'uscita della chiglia che migliora l'angolo di bolina a scapito della resistenza), i movimenti dello scafo e la posizione di un'altra barca che fungeva da lepre. Queste informazioni venivano trasmesse alla barca appoggio, dove un calcolatore (un MicroVax II della Digital Equipment Corporation) immagazzinava i dati che successivamente venivano analizzati a terra da un altro MicroVax II. Le informazioni ottenute nel corso di queste prove sono state di grande importanza per convalidare i risultati forniti dal VPP e fare esperienza su barche vere in condizioni reali.

Il nostro incontro con la tecnologia ci ha portati a considerare altri fattori non direttamente inerenti alla barca. Le regate hanno sempre una componente casuale perché il comportamento delle due imbarcazioni in gara dipende dalla velocità del vento e dalle condizioni del mare, che variano casualmente ogni giorno. I risultati ottenuti con il VPP non sono quindi, di per sé, conclusivi, anzi, possono essere fuorvianti nel confrontare su una serie di regate i meriti di due imbarcazioni. Abbiamo pure scoperto con sorpresa, anche se poi la cosa è risultata abbastanza ovvia, che non é possibile stabilire quale barca sia la migliore per una determinata serie di regate, senza conoscere le caratteristiche degli altri concorrenti. Quindi è stato necessario venire a sapere tutto ciò che era possibile sulle barche che probabilmente avremmo incontrato e sviluppare metodi basati sulla teoria della probabilità e la teoria dei giochi per valutare come il nostro candidato si sarebbe comportato in regata contro una ipotetica flotta di sfidanti e di difensori.

A questo fine abbiamo sviluppato due programmi per simulare le regate fra due barche. Il primo si basa su un semplice modello probabilistico e prende in considerazione diversi fattori: il distacco fra i contendenti in funzione della velocità del vento; la distribuzione di probabilità di quest'ultima nel mese della regata; una rappresentazione di fattori casuali quali la variabilità del vento, le condizioni del mare, piccoli errori o incidenti e, infine, un possibile abbuono per una delle due barche, che premia la miglior scelta delle vele, la tattica più intelligente o il maggiore allenamento.

Il secondo programma è molto più elaborato. Oltre ai dati del programma precedente, considera pure la storia del vento per una giornata tipica; le differenze nelle lunghezze dei bordi in funzione della velocità del vento, le interazioni fra le due barche quando sono vicine fra di loro e il risultato incerto di ingaggi molto serrati. Entrambi i programmi forniscono in uscita una probabilità: per esempio che, con i venti di novembre al largo di Perth, la barca *A* può battere la barca *B* nel 59 per cento delle regate.

Abbiamo deciso di impiegare la teoria dei giochi perché essa fornisce strategie razionali per prendere decisioni in una competizione. Siamo riusciti ad applicare la teoria per due persone a somma nulla per aiutare nella scelta della barca ottimale per le nostre condizioni di regata (somma nulla significa che la vittoria di un contendente comporta la perdita per il secondo). Poniamo che il rosso scelga una barca di 46 piedi di lunghezza (un piede corrisponde a circa 30 centimetri), mentre il blu (noi) ne scelga una da 48 piedi. Senza dare alcun vantaggio a nessuna delle due barche, potremmo vincere solo il 28 per cento delle regate in novembre e il 72 per cento in gennaio. Quale dovrà essere la lunghezza della barca da scegliere per avere il massimo delle possibilità se una delle due squadre ha un vantaggio perché è più abile?

Le possibilità possono venire organizzate in matrici. La matrice dei vantaggi e svantaggi nella quale viene dato al blu un vantaggio di 30 secondi mostra che la sua miglior scelta rimane una barca di 48 piedi, mentre il rosso può sceglierne una di 45 o di 47 piedi e poi sperare che le condizioni lo favoriscano. Anche così il rosso avrebbe solamente il 34 per cento di probabilità di vittoria in tutte le regate.

I finalisti della sfida della Coppa America nel 1987 sono stati *Stars & Stripes*, qui in testa, e l'australiana *Kookaburra III*. *Stars & Stripes* ha raggiunto la finale eliminando 12 sfidanti degli Stati Uniti e di cinque altre nazioni. *Kookaburra III* ha superato cinque imbarcazioni australiane. Nella serie finale di sette regate, *Stars & Stripes* ha vinto nettamente con quattro regate.

Tutte le tecnologie che abbiamo descritto hanno concorso, in una strategia comune, alla progettazione di *Stars & Stripes*. La sua stessa esistenza è dovuta ai programmi di previsione della velocità e di simulazione delle regate perché, senza le loro chiare indicazioni, sicuramente non sarebbe stata costruita. I nostri precedenti 12 metri, pure denominati in un primo momento *Stars & Stripes*, ma successivamente designati *'85* e *'86* in base al loro anno di costruzione, avevano dato prova di una potenza, una stabilità e una velocità senza precedenti con venti forti. In particolare, *Stars & Stripes '85* aveva dimostrato di essere leggermente più veloce nella maggior parte delle situazioni e il suo equipaggio aveva preso una grandissima confidenza con quella barca.

Ma, fra la fine dell'85 e l'inizio dell'86, mentre la squadra di Sail America stava allenandosi alle Hawaii, i candidati australiani alla difesa della coppa e molti degli sfidanti si allenavano a Perth, impegnandosi nel Campionato mondiale dei 12 metri. Osservazioni accurate, fra cui anche lo studio fotogrammetrico degli scafi, ci dimostrò che queste barche erano decisamente più piccole di quanto noi avevamo stimato giusto per le condizioni estive. I risultati della simulazione delle regate indicarono che, per quanto *Stars & Stripes '85* avesse elevate probabilità di battere tutti i competitori conosciuti vincendo le regate in gennaio e in febbraio, le sue probabilità di sopravvivere alle eliminatorie contro una flotta di sfidanti più corti di due o tre piedi sarebbero state assai limitate.

Di fronte a queste previsioni, Sail America non aveva altra possibilità che costruire un'altra barca, abbastanza piccola da poter competere con successo nelle eliminatorie. Il nuovo progetto avrebbe pure potuto avvantaggiarsi delle nuove forme dello scafo sviluppate in base alla teoria del bastimento sottile, già provate in vasca, e dei progressi nell'ottimizzazione della forma della chiglia. La teoria dei giochi ha suggerito di scegliere uno scafo leggermente più lungo di quelli del resto della flotta. Dopo le eliminatorie si sarebbero potute aumentare l'apertura delle alette e la zavorra (per aumentare la lunghezza al galleggiamento) e si sarebbe potuto accentuare la forma a bulbo della chiglia, abbassando il baricentro. Così modificata per sopportare i forti venti estivi, la barca sarebbe stata difficile da battere.

Questa barca venne costruita in tempo record e provata alle Hawaii abbastanza a lungo da riuscire a dimostrare la sua superiorità rispetto alle barche precedenti. Il successo della nostra strategia fa ora parte della storia dello yachting.

Collisioni tra protoni in rotazione

Sorprendenti risultati sperimentali, che dimostrano l'importanza del verso dello spin nel determinare l'esito di una collisione, potrebbero indurre a rivedere la descrizione teorica oggi accettata della struttura del protone

di Alan D. Krisch

Sembra che tutte le particelle fondamentali che costituiscono la materia - protoni, neutroni ed elettroni - ruotino come trottole. La rotazione è una proprietà quantomeccanica fondamentale; ogni particella possiede una quantità definita di spin, o momento angolare di spin, proprio come possiede una massa definita e una carica elettrica definita. Quando due particelle che ruotano entrano in collisione, il verso del loro spin può influenzare la loro successiva diffusione, come l'«effetto» impresso alle palle da biliardo può alterare la loro traiettoria dopo un urto.

Un abile giocatore con una conoscenza intuitiva delle proprietà di una palla da biliardo può eseguire colpi difficili, ma prevedibili regolando la rotazione (spin) e la velocità della palla. In una serie di esperimenti con acceleratori di particelle, i miei colleghi e io abbiamo ugualmente fatto variare lo spin e l'energia dei protoni in collisione. Ma al contrario del giocatore di biliardo, non abbiamo potuto prevedere gli effetti dei nostri «colpi», perché molte proprietà del protone sono ancora misteriose. In effetti, abbiamo osservato comportamenti inattesi e spesso sorprendenti che sembrano in disaccordo con l'attuale teoria sulla struttura del protone e sulle forze che si esercitano al suo interno, la cromodinamica quantistica (QCD).

La ricerca iniziò nel 1973, con lo Zero Gradient Synchrotron (ZGS), dell'Argonne National Laboratory, dove il mio gruppo di ricerca dell'Università del Michigan fece diffondere fasci di protoni polarizzati da bersagli nei quali i protoni erano anch'essi polarizzati, ovvero tutti con la stessa direzione di spin (si veda l'articolo *Lo spin del protone* di Alan D. Krisch in «Le Scienze» n. 131, luglio 1979). I risultati dimostrarono che lo spin ha un ruolo significativo nelle interazioni di alta energia tra protoni: violente collisioni protone-protone si verificavano principalmente quando il fascio e il bersaglio erano polarizzati nella stessa direzione. Quando i protoni del fascio e quelli del bersaglio avevano spin opposti, sembrava spesso che si attraversassero senza interagire.

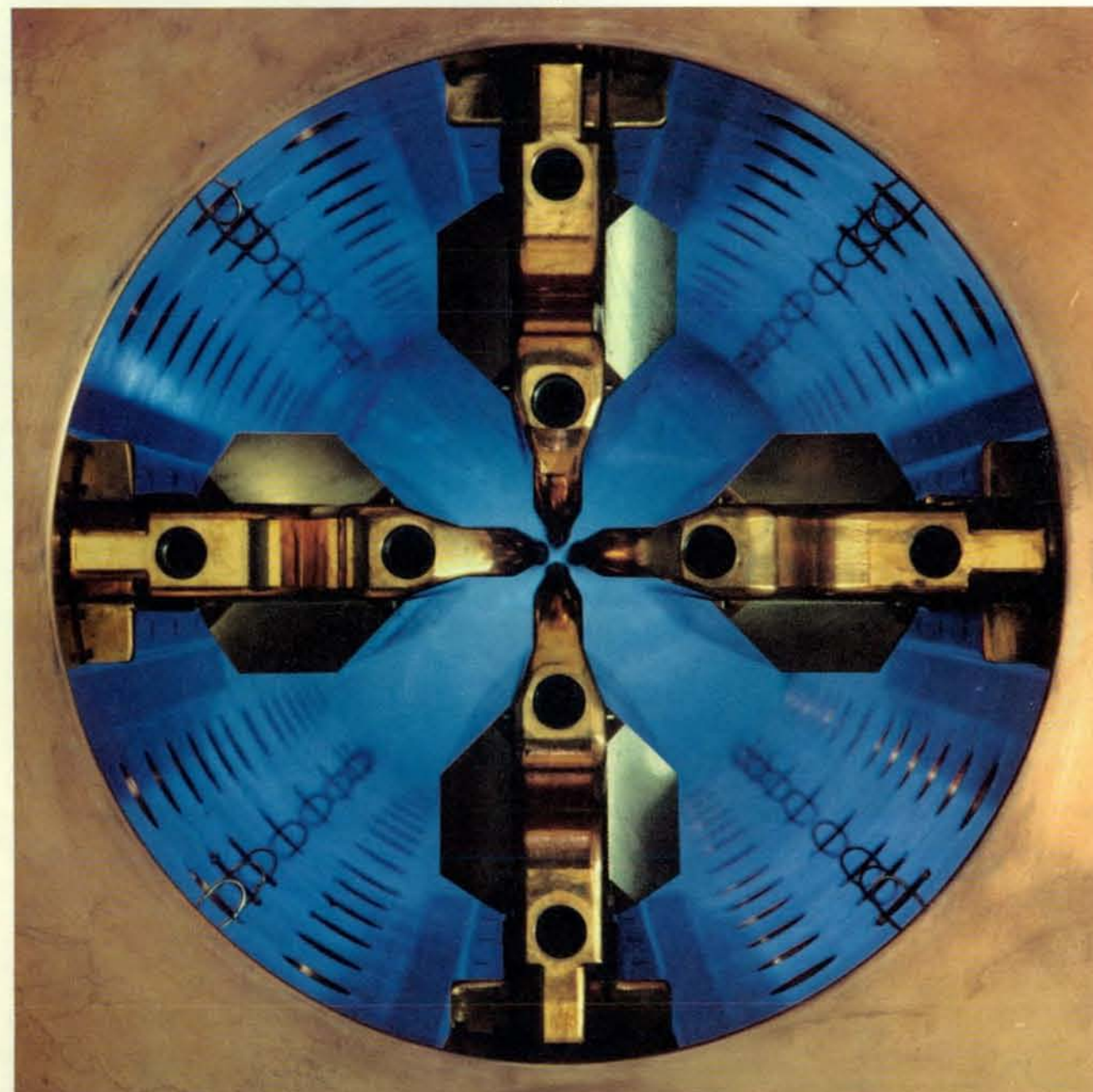
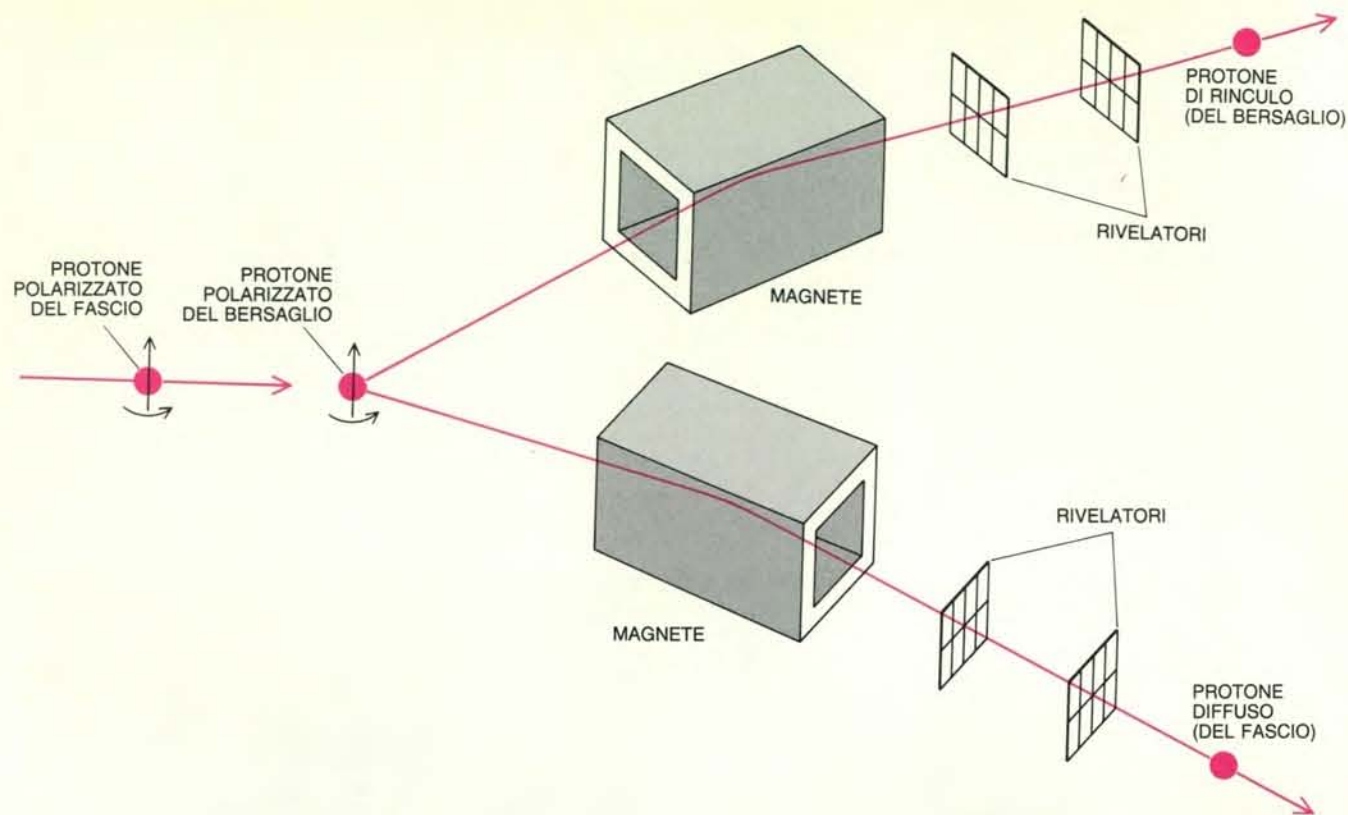
Utilizzando lo ZGS, dovevamo limitarci a energie sperimentali di circa 13 GeV (miliardi di elettronvolt). Dopo sei anni di laborioso adattamento di un altro acceleratore, l'Alternating Gradient Synchrotron (AGS), del Brookhaven National Laboratory, abbiamo ora potuto indagare gli effetti dello spin del protone a energie molto più alte: fino a 18,5 GeV, quando sia il fascio sia il bersaglio erano polarizzati, e a 28 GeV, con il solo bersaglio polarizzato. Le sorprese si sono moltiplicate: all'aumentare dell'energia di collisione dei protoni, gli effetti dello spin sembrano oscillare. Alle energie più alte raggiunte, gli effetti dello spin sono particolarmente accentuati.

Da un certo punto di vista, è facile comprendere la nostra sorpresa. L'energia associata allo spin di un protone è costante, quindi il ruolo giocato dallo spin dovrebbe diminuire all'aumentare dell'energia di collisione. A energie sufficientemente alte, dovrebbe essere indifferente che i due protoni in collisione stiano ruotando nello stesso verso o in verso opposto. Il fatto che le direzioni di spin introducano una grande differenza fa pensare che le nostre conoscenze sulle interazioni reciproche dei protoni siano incomplete. Gli esperimenti rimettono in questione persino il modello attualmente accettato della struttura interna del protone, secondo il quale esso è formato da tre costituenti più piccoli, i quark, tenuti insieme dalla forza nucleare forte (la forza descritta dalla QCD).

I nostri nuovi risultati non sono certo il primo caso in cui il fenomeno dello spin ha sorpreso e confuso i fisici. Quando Samuel A. Goudsmit e George E. Uhlenbeck proposero per la prima volta, nel 1925, che gli elettroni dovessero possedere uno spin, Wolfgang Pauli affermò che l'idea era folle. Però, alla fine degli anni venti e negli anni trenta, lo spin divenne un elemento importante nello sviluppo della meccanica quantistica e della fisica atomica, in particolare nei lavori di P. A. M. Dirac.

Durante gli anni quaranta, i fisici nucleari erano convinti che gli effetti dello spin venissero sentiti solo in collisioni atomiche di bassa energia e che non avrebbero potuto essere significativi in collisioni nucleari con energie di parecchi milioni di elettronvolt. Successivamente, Charles L. Oxley dell'Università di Rochester e colleghi ottennero la prova che forti effetti di spin avevano luogo anche a energie di parecchie centinaia di milioni di elettronvolt. La sorpresa e l'eccitazione che seguirono questo annuncio sono documentate in un articolo di Enrico Fermi pubblicato nel 1954.

La maggior parte dei fisici delle alte energie riteneva che lo spin non fosse importante in collisioni tra particelle elementari a energie di miliardi di elettronvolt. Per anni, questa idea non fu verificata, se non in una serie di difficili esperimenti eseguiti, tra gli altri, da Owen Chamberlain ed Emilio Segré dell'Università della California a Berkeley. Poi, al termine degli anni cinquanta, Anatole Abragam del Collège de France e Carson D. Jeffries di Berkeley suggerirono di produrre bersagli di protoni polarizzati. La tecnica, che ha avuto pieno successo, consiste nel polarizzare, mediante basse temperature e un intenso campo magnetico, gli spin di alcuni elettroni in compresse congelate di materiale bersaglio; il campo magnetico fa «allineare» gli spin degli elettroni. Il campione viene esposto poi a radiazione di microonde e l'allineamento degli spin degli elettroni viene trasferito ai protoni vicini, facendoli ruotare tutti nello stesso verso. Gli esperimenti che impiegavano bersagli di protoni polarizzati, eseguiti negli anni sessanta e settanta a Berkeley, al CERN (Organizzazione europea per la ricerca nucleare) e all'Argonne National Laboratory rivelarono piccoli, ma interessanti effetti di spin in collisioni di alta energia. Nonostante ciò, la maggior parte dei fisici rimaneva convinta che lo spin non fosse molto importante e lo sarebbe stato sempre meno a energie superiori.

Nel 1973, il mio gruppo di ricerca, utilizzando lo ZGS, inaugurò un metodo diverso, polarizzando sia il fascio sia il bersaglio. La polarizzazione del bersaglio, raffreddato a mezzo grado sopra lo zero assoluto con una miscela di elio 3 ed elio 4 liquidi, era assicurata da uno speciale magnete, che generava un forte campo di 25 000 gauss. Nello stesso tempo un complesso sistema di magneti conservava la polarizzazione del fascio di protoni, mentre veniva accelerato. Facendo diffondere i protoni del fascio polarizzato per mezzo dei protoni del bersaglio polarizzato, eseguimmo le prime misure di collisioni elastiche tra protoni rotanti in direzioni note. Una collisione è elastica se tutta l'energia del protone incidente viene portata via dai due protoni di rinculo, mentre in una collisione anelastica una parte dell'energia viene utilizzata per creare nuove particelle.

I protoni, nell'Alternating Gradient Synchrotron (AGS) del Brookhaven National Laboratory, vengono accelerati dalla differenza di potenziale creata da un quadrupolo a radiofrequenza (RFQ). I poli dell'RFQ, due positivi e due negativi, sono disposti in modo tale che poli identici risultino affacciati. Nel dispositivo dell'illustrazione, lungo 1,5 metri e con un diametro di 30 centimetri, le polarità si invertono alla frequenza di 200 milioni di volte al secondo. I protoni che entrano nell'RFQ sono già stati polarizzati: i loro spin hanno tutti lo stesso verso. L'RFQ conserva la polarizzazione e, nel contempo, incrementa l'energia dei protoni da 20 a 760 keV.

Un fascio polarizzato di protoni viene diffuso da un bersaglio fisso di protoni polarizzati. Rivelatori di particelle registrano gli angoli secondo i quali avviene la diffusione dei protoni; l'angolo di deflessione nei magneti di curvatura indica l'energia dei protoni dopo l'urto.

Abbiamo studiato soltanto collisioni elastiche poiché esse ci sembrano, nello stesso tempo, semplici e fondamentali.

Abbiamo studiato la diffusione di protoni in quattro stati iniziali di spin. Nel primo gli spin dei protoni del fascio e di quelli del bersaglio erano entrambi verso l'alto (cioè, se le dita della mano destra si avvolgono nel verso di rotazione, il pollice è diretto verso l'alto). Chiamerò questo stato su-su, dove il primo «su» si riferisce alla direzione dello spin dei protoni del fascio e il secondo a quella dei protoni del bersaglio. Secondo tale convenzione, i restanti stati iniziali vengono chiamati giù-giù, su-giù e giù-su. La direzione dello spin dei protoni del fascio veniva invertita ogni tre secondi circa, mentre la direzione dello spin dei protoni del bersaglio veniva invertita a intervalli di qualche ora.

Per sondare veramente a fondo l'interno del protone, abbiamo cercato di rendere le collisioni il più possibile violente. Speravamo di trovare nuovi indizi sulla natura dei quark, i costituenti del protone, che sembra non possano condurre vita indipendente all'esterno di esso. La teoria dei quark, messa a punto da Murray Gell-Mann del California Institute of Technology, è riuscita a spiegare, con ottimi risultati, la massa delle numerose particelle a vita breve create nella collisione di protoni. D'altra parte, l'estensione della teoria alla diffusione di particelle - la cromodinamica quantistica (QCD) - ha fornito poche previsioni verificabili. La QCD è una teoria molto flessibile che riesce facilmente a spiegare la maggior parte dei nuovi dati accumulati sulla diffusione; per quanto mi riguarda, essendo particolarmente attento agli aspetti formali di una teoria, sono rimasto impressionato non dall'adattabilità, ma dal potere previsionale.

Devo anche confessare un certo imbarazzo di fronte all'idea che i quark possano esistere come particelle all'interno e non all'esterno di un protone. Le brillanti e attraenti ipotesi che sono state avanzate, nell'ambito della QCD, per spiegare l'apparente confinamento dei quark potranno rivelarsi corrette, e forse io finirò per cambiare la mia concezione di vecchio stampo secondo cui una particella deve essere un oggetto ben definito. Penso, però, che un concetto semplice non dovrebbe essere abbandonato a favore di uno più complesso, a meno di schiaccianti prove sperimentali.

Le migliori collisioni, per sondare il più a fondo possibile le forze di spin delle particelle costituenti il protone, sono quelle in cui i due protoni, dopo l'urto, rinculano esattamente ad angolo retto rispetto alla direzione iniziale di moto. Tali collisioni sono le più violente perché trasferiscono la massima frazione dell'energia del protone incidente. Nel nostro esperimento finale allo ZGS abbiamo studiato collisioni perpendicolari variando l'energia dei protoni incidenti tra 4 e 13 GeV; per ogni valore di energia, abbiamo osservato ciò che accadeva quando i protoni ruotavano nello stesso verso e quando ruotavano in versi opposti.

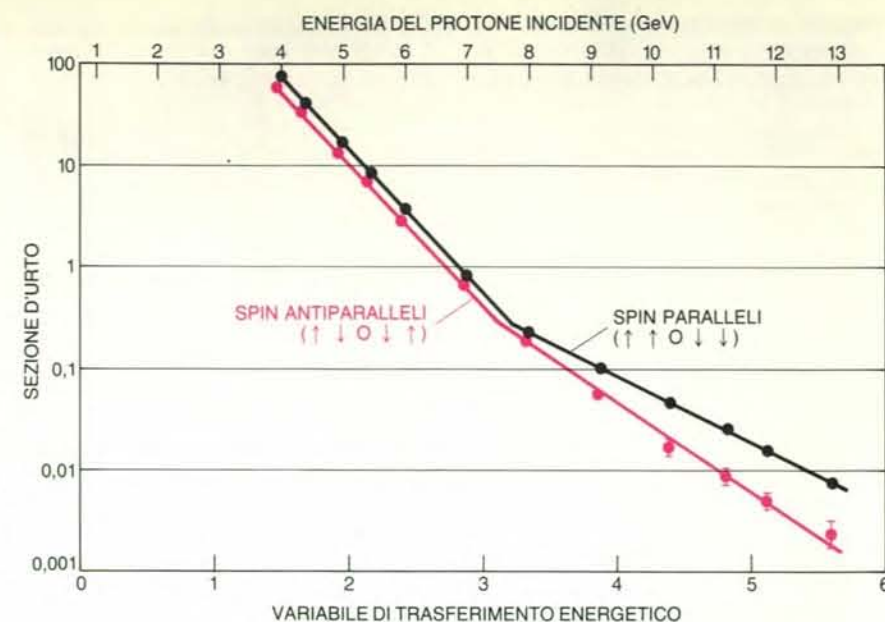
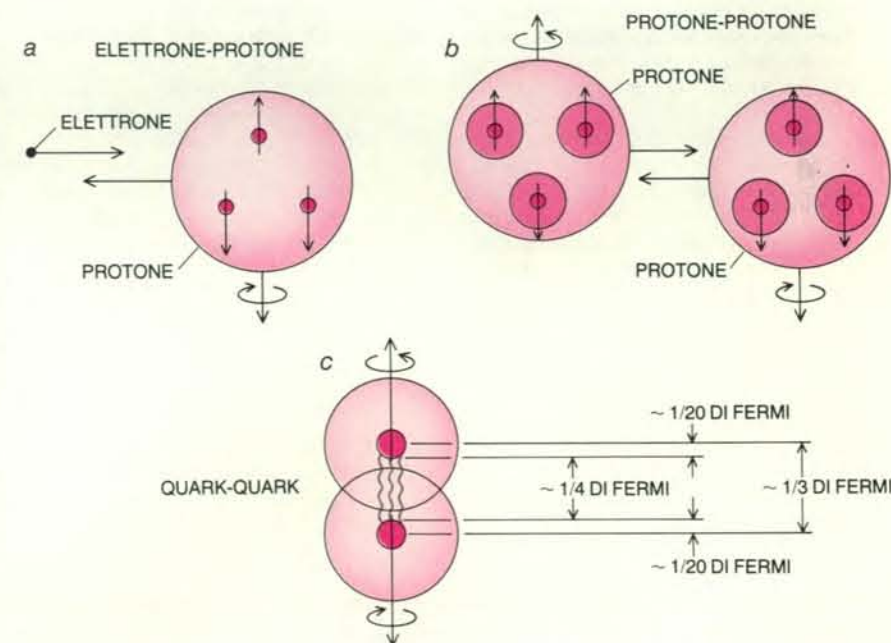
In particolare, abbiamo misurato separatamente la probabilità di una collisione elastica esattamente perpendicolare per protoni con spin paralleli (su-su o giù-giù) e per protoni con spin antiparalleli (su-giù o giù-su). Tale probabilità di collisione è detta sezione d'urto per la diffusione elastica protone-protone a 90 gradi. La sezione d'urto, che si può interpretare come la dimensione efficace dei protoni interagenti, viene rappresentata normalmente in funzione dell'energia della particella incidente.

Il nostro esperimento ha rivelato due importanti indizi sui costituenti del protone. Il primo è il fatto che la sezione d'urto diminuisce rapidamente all'aumentare dell'energia di collisione fino a 8 GeV, mentre al di sopra di 8 GeV diminuisce solo gradualmente. I dati possono fornire informazioni sulla struttura del protone mediante analisi di Fourier. La rapida diminuzione iniziale della sezione d'urto indica che a bassa energia i protoni si limitano a rimbalzare uno sull'altro; sembra quasi che lo strato esterno del protone sia relativamente morbido e grande, con un raggio di circa un fermi, ovvero 10^{-15} metri. La successiva diminuzione graduale della sezione d'urto fa pensare che le collisioni di energia maggiore interessino oggetti rigidi all'interno del protone, con un raggio di circa un terzo di fermi. Tale risultato è in accordo col modello a tre quark della QCD.

Il secondo indizio sui costituenti del protone è invece problematico per la QCD. Esso si basa sull'osservazione che, a energie superiori a 8 GeV, la sezione d'urto diminuisce più rapidamente quando gli spin dei protoni sono antiparalleli invece che paralleli. In altri termini, i protoni hanno, per qualche ragione, una maggiore probabilità di collidere violentemente quando i loro spin sono paralleli. A 13 GeV la sezione d'urto (probabilità di collisione) è quattro volte maggiore quando gli spin sono paralleli che non quando sono antiparalleli. Anche se non sappiamo esattamente quale sia la causa di questo comportamento strano e del tutto inatteso, esso non appare di buon auspicio per la QCD.

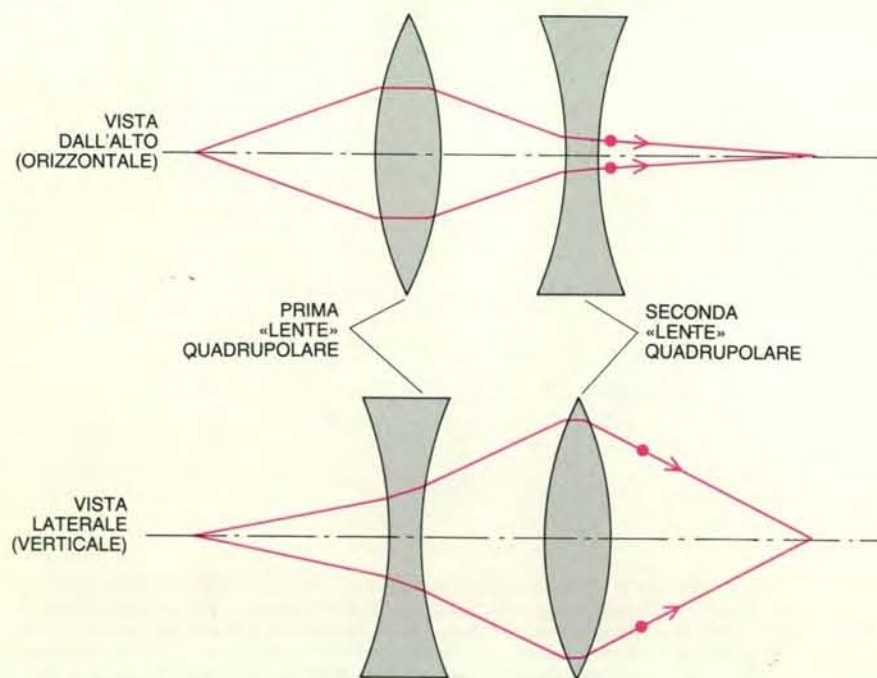
Secondo la teoria, due dei tre quark in ogni protone dovrebbero ruotare nello stesso verso del protone, mentre il terzo quark dovrebbe ruotare nel verso opposto. Perciò, indipendentemente dal fatto che gli spin dei due protoni in collisione siano paralleli o antiparalleli, gli urti coinvolgono sempre qualche coppia di quark i cui spin sono paralleli e qualche coppia i cui spin sono antiparalleli. I calcoli condotti nell'ambito della QCD ammettono una differenza di un fattore due tra le sezioni d'urto a spin paralleli e quelle a spin antiparalleli, ma una differenza di un fattore quattro non può facilmente conciliarsi con la teoria.

Che cosa significa tale differenza? Forse la QCD vale soltanto per collisioni più energetiche e più violente di quelle da noi studiate? Purtroppo non vi è accordo fra i teorici sul possibile inizio dell'intervallo di validità della QCD. Dal momento che il rapporto fra le sezioni d'urto sembra aumentare quando le collisioni diventano più violente ed energetiche, i nostri dati non sono certo di sostegno alla teoria. Una spiegazione è che i costituenti «visti» negli esperimenti di diffusione - anche se in qualche modo correlati ai quark che giustificano la massa delle particelle - non siano esattamente i tre quark rotanti previsti dalla QCD. Forse ogni protone contiene altri costituenti, o forse, in alternativa, i tre costituenti cessano di comportarsi come quark indipendenti quando i protoni entrano in collisione.

Nei nostri esperimenti di diffusione a 90 gradi abbiamo raggiunto il limite energetico di 13 GeV dello ZGS. L'estensione degli esperimenti a collisioni ancora più violente richiedeva un fascio di protoni polarizzati di energia più elevata. Alla fine degli anni settanta, un gruppo di circa 80 scienziati, ingegneri, tecnici e studenti dell'Argonne National Laboratory, di Brookhaven, dell'Università del Michigan, della Rice University e della Yale University cominciò a modificare l'AGS di Brookhaven per ac-

La diffusione dei protoni alle alte energie dipende dallo spin. La sezione d'urto a 90 gradi, ossia la probabilità di diffusione secondo tale angolo, è rappresentata graficamente in funzione della variabile di trasferimento energetico delle collisioni. (Dato un angolo di diffusione, la variabile di trasferimento energetico è proporzionale all'energia del protone incidente.) Nel grafico sono rappresentati due gruppi di dati, uno in cui gli spin dei protoni incidenti e di quelli del bersaglio sono paralleli e uno in cui gli spin sono antiparalleli. A bassi valori di energia, le sezioni d'urto a spin paralleli e antiparalleli sono identiche, come è previsto dalla teoria più accreditata della struttura e delle proprietà del protone, la cromodinamica quantistica (QCD). A energie più elevate, però, per ragioni non ancora chiarite, le sezioni d'urto diventano notevolmente diverse.

La descrizione teorica del protone secondo la QCD postula che la particella sia formata da tre costituenti, i quark, legati dalla forza forte. Esperimenti di diffusione di elettroni puntiformi da parte di protoni in quiete, compiuti allo Stanford Linear Accelerator Center (*a*), hanno permesso a ricercatori del Massachusetts Institute of Technology e dello SLAC di dimostrare che i costituenti del protone hanno un raggio del valore di meno di 1/20 di fermi. (Un fermi è uguale a 10^{-15} metri, pari circa al raggio del protone.) I dati ottenuti negli esperimenti dell'autore, in cui vengono diffusi protoni (*b*), indicano che i costituenti devono avere un raggio di circa 1/3 di fermi. Combinando i due risultati, si ottiene un modello interessante e forse corretto dei costituenti del protone (*c*), secondo cui i dati dello SLAC indicano le dimensioni del campo elettrico generato dai costituenti. Ogni costituente può avere perciò un nucleo con un raggio di circa 1/20 di fermi e un forte campo di forza che si estende all'esterno per circa 1/4 di fermi. Quando due quark si colpiscono in una collisione, le loro dimensioni efficaci sono di 1/3 di fermi.

celerare i protoni fino a circa 20 GeV. Si trattava di un compito difficile poiché, durante il ciclo di accelerazione di mezzo secondo, i protoni, circolando ripetutamente entro il sincrotrone, percorrono circa 150 000 chilometri all'interno di intensi campi magnetici, che possono facilmente distruggerne la polarizzazione, facendoli ruotare in direzioni differenti. I campi magnetici dell'AGS sono molto più intensi di quelli dello ZGS e di conseguenza il problema è più grave. In effetti, un comitato della US Atomic Energy Commission, di cui facevo parte, concluse nel 1974 che «motivi tecnici potrebbero impedire di accelerare protoni polarizzati in acceleratori [del tipo AGS]».

L'inconveniente della depolarizzazione si rivelò un ostacolo difficile, ma infine riuscimmo a superarlo. Per raggiungere l'energia massima di 22 GeV che ci eravamo proposti, fummo costretti ad affrontare il problema delle «risonanze» depolarizzanti, che si manifestavano a 45 diverse energie nel ciclo di accelerazione. Navigare nel mare delle risonanze depolarizzanti costituì uno sforzo formidabile, che richiese l'impiego di sofisticati strumenti, i polarimetri, per valutare la percentuale di protoni rotanti in ciascuna direzione. Tale percentuale ci permise di determinare se eravamo sulla strada giusta nel regolare i magneti appositamente progettati per consentire ai protoni di superare rapidamente le energie di risonanza. Il gruppo che si occupava del problema, guidato da Lazarus G. Ratner di Brookhaven e da me, dovette lavorare a tempo pieno per sei settimane per aver ragione delle 45 risonanze depolarizzanti. Nel gennaio 1986, dopo sei anni di intensi sforzi per tutte le modifiche, riuscimmo ad accelerare un fascio polarizzato a 18,5 GeV, un nuovo record mondiale. Gli esperimenti con questo apparecchio straordinario iniziarono immediatamente e in poche settimane raggiungemmo i 22 GeV.

Un sincrotrone utilizza sia campi elettrici sia campi magnetici per accelerare i protoni (che possiedono una carica elettrica) fino a energie di molti miliardi di elettronvolt. I protoni sono costretti a muoversi in un fascio circolare all'interno di una camera a vuoto a forma di ciambella da un campo magnetico perpendicolare al piano dell'anello. A ogni giro dell'anello, il protone riceve un incremento di energia di circa 100 keV nella cavità di accelerazione.

Il principio della focalizzazione forte permette di confinare i protoni in moto entro un sincrotrone in una regione a forma di ciambella di piccola sezione e di ridurre così le dimensioni dei magneti necessari per curvare il fascio di protoni, disponendo uno dopo l'altro due magneti quadrupolari opposti. Il primo focalizza il fascio di protoni orizzontalmente, ma lo fa divergere verticalmente, mentre il secondo focalizza il fascio verticalmente, ma lo fa divergere orizzontalmente. Regolando opportunamente le intensità dei magneti, è possibile focalizzare il fascio in entrambe le direzioni.

Per spiegare le cause della depolarizzazione e in che modo siamo riusciti a superarla, devo prima descrivere come funziona un sincrotrone. Come il ciclotrone, esso utilizza sia campi elettrici sia campi magnetici per accelerare i protoni (che hanno una carica elettrica) ad alte energie. I protoni sono costretti a muoversi in cerchio entro un anello a forma di ciambella per opera di un campo magnetico perpendicolare al piano dell'anello. Tale campo di curvatura verticale è efficace nel provocare la depolarizzazione: se l'asse di spin di un protone è orizzontale, il campo può farlo ruotare più volte durante ogni giro nell'anello, rendendo quasi impossibile conservare l'allineamento di tutti gli assi di spin in un fascio di protoni.

Ogni volta che un protone compie un giro nella ciambella, riceve un incremento di energia di circa 100 000 elettronvolt dalla differenza di potenziale creata nella cavità di accelerazione. Durante il ciclo di accelerazione dell'AGS, ciascun protone percorre gli 800 metri di circonferenza della ciambella circa 200 000 volte, acquistando un'energia totale di circa 20 GeV alla fine del ciclo. Ogni giro nell'anello dell'AGS dura 0,0000025 secondi (2,5 microsecondi), e quindi l'intero ciclo richiede circa mezzo secondo.

È difficile mantenere i protoni all'interno della ciambella per l'intero ciclo di accelerazione, dato che ogni particella percorre circa 160 000 chilometri all'interno di un tubo che ha un diametro di soli otto centimetri circa. Anche il più abile tiratore difficilmente colpirebbe un bersaglio di otto centimetri da una distanza di 160 000 chilometri. A questo scopo, si utilizzano campi magnetici che focalizzano i protoni e impediscono loro di sfuggire. I campi magnetici possono focalizzare particelle cariche nello stesso modo in cui una lente di ingrandimento focalizza i raggi di luce.

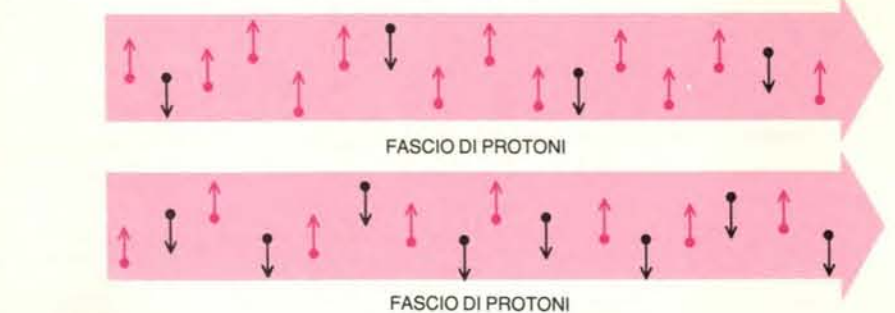
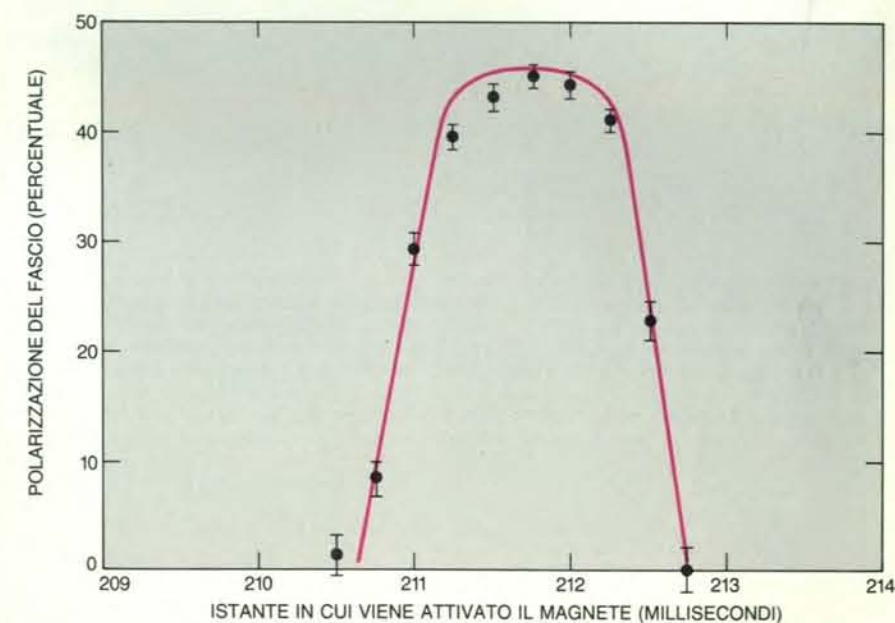
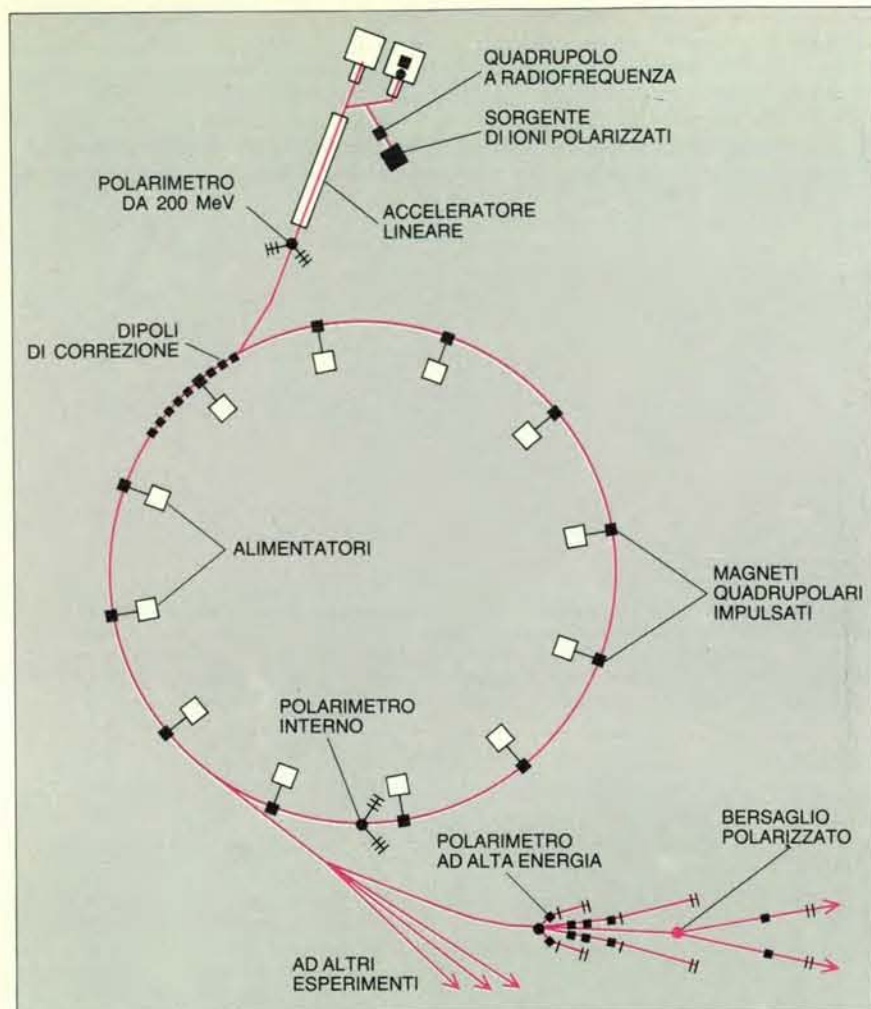
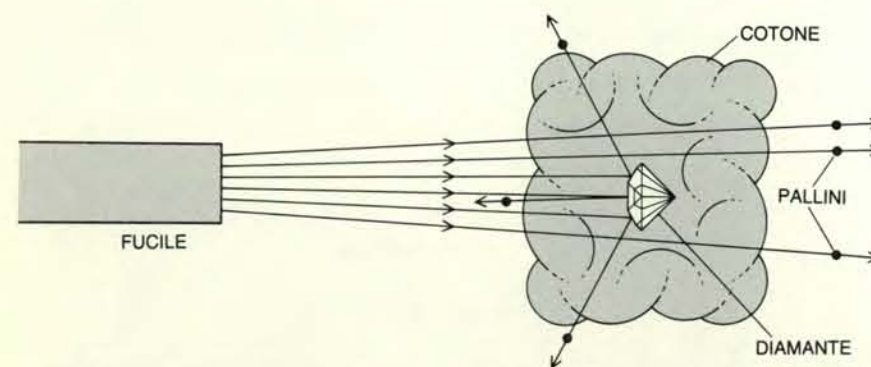
I primi sincrotroni, quali lo ZGS, usavano campi magnetici di focalizzazione piuttosto deboli e furono perciò chiamati sincrotroni a focalizzazione debole. Il fascio di protoni, scarsamente focalizzato, divergeva e diventava molto largo e quindi la camera a vuoto a ciambella contenente il fascio doveva essere di grande sezione. La camera a vuoto dello ZGS era larga circa 80 centimetri e alta circa 15. Le grandi dimensioni avevano alcuni vantaggi; ricordo, nel 1965, di aver introdotto la mano nella ciambella per cercare di fissare un bersaglio rotto, senza dover smontare il magnete da 600 tonnellate che lo circondava. Purtroppo, una camera a vuoto grande deve essere circondata da magneti ancor più grandi, il cui costo, per gli acceleratori a energia più elevata, è proibitivo. I magneti dello ZGS erano larghi circa 2,5 metri e alti circa 1,5 metri e pesavano 30 tonnellate per metro di lunghezza. Lo ZGS, la cui costruzione fu iniziata nel 1955, fu l'ultimo sincrotrone a focalizzazione debole costruito.

Per raggiungere energie più elevate era necessaria un'idea migliore. Negli anni cinquanta Nicholas C. Christofilos, Ernest D. Courant, Stanley Livingston e Hartland S. Snyder inventarono il principio della focalizzazione forte, che impiega magneti quadrupolari, ognuno dei quali è formato da due poli magnetici nord e due poli magnetici sud disposti alternativamente ai vertici di un quadrato. Quando un fascio di protoni attraversa il quadrato, i protoni che deviano dall'asse del fascio sono sottoposti a una forza focalizzante sempre maggiore che li riavvicina all'asse. Purtroppo un solo magnete quadrupolare non può focalizzare i protoni in direzione orizzontale e in quella verticale. Se il magnete focalizza un fascio di protoni in una direzione, le particelle divergono nell'altra direzione e sfuggono dalla camera a vuoto.

Il segreto della focalizzazione forte è di disporre uno dopo l'altro magneti quadrupolari opposti. Il primo focalizza il fascio orizzontalmente, ma lo fa divergere verticalmente, mentre il secondo lo focalizza verticalmente, ma lo fa divergere orizzontalmente. A prima vista, si potrebbe pensare che focalizzazione e divergenza si annullino a vicenda senza ottenere alcun risultato. Invece, regolando adeguatamente le intensità dei due magneti quadrupolari, si può focalizzare un fascio in entrambe le direzioni. Tutti gli acceleratori costruiti dopo il 1960 utilizzano la focalizzazione forte.

Le coppie di magneti quadrupolari costringono i protoni a oscillare intorno all'asse centrale dell'anello del sincrotrone, in onde chiamate oscillazioni di betatrone. Le oscillazioni di betatrone verticali fanno viaggiare i protoni attraverso campi magnetici orizzontali, che depolarizzano i protoni con asse di spin verticale. I campi sono deboli, ma, se il protone ne incontra un numero sufficiente, il suo spin può variare. Più creste e ventri vi sono nelle oscillazioni di betatrone, più spesso il protone attraverserà un campo magnetico orizzontale in ogni giro dell'anello.

Fortunatamente, i problemi di depolarizzazione associati alle oscillazioni di betatrone e ai campi magnetici orizzon-

Le oscillazioni verticali di betatrone in un fascio di protoni sono moti ondulatori verticali indotti dalle coppie di magneti per la focalizzazione forte posti intorno alla circonferenza del sincrotrone. Tali moti mantengono i protoni lontani dalle pareti della camera a vuoto, mentre percorrono l'anello. Poiché le oscillazioni fanno attraversare ai protoni campi magnetici orizzontali, esse possono distruggere la polarizzazione del fascio, impartendo direzioni casuali agli spin.

La polarizzazione dei protoni nell'AGS viene conservata attivando speciali magneti quadrupolari impulsati in momenti accuratamente calcolati durante il ciclo di accelerazione dei protoni. Il gruppo dell'autore ha ottenuto un fascio polarizzato per il 50 per cento (*in alto*): il 75 per cento dei protoni aveva un verso di spin e il 25 per cento aveva il verso opposto (*al centro*). Senza un'opportuna scelta dei tempi degli impulsi, la percentuale di polarizzazione cadeva a zero, ossia circa metà dei protoni mostrava un verso di spin e metà mostrava l'altro verso (*in basso*).

Il sincrotrone a gradiente alternato è in grado di accelerare i protoni polarizzati fino a un'energia di 22 GeV (miliardi di elettronvolt). Una sorgente di ioni polarizzati genera protoni che presentano spin per la maggior parte paralleli. Il quadrupolo a radiofrequenza crea una differenza di potenziale che accelera i protoni fino a 760 keV prima di immetterli in un acceleratore lineare, che, a sua volta, li porta fino a 200 MeV (milioni di elettronvolt). Dopo essere usciti dall'acceleratore lineare, i protoni vengono introdotti nell'anello principale del sincrotrone. I magneti quadrupolari impulsati e i dipoli di correzione permettono di superare i problemi di depolarizzazione. Il grado di polarizzazione durante il ciclo è controllato da tre polarimetri.

Gli esperimenti di diffusione permettono di sondare la struttura del protone analogamente a come un immaginario ispettore di dogana potrebbe cercare un diamante nascosto in una grossa balla di cotone. L'ispettore spara un colpo di fucile nella balla e osserva il numero di pallini che diffondono in ogni direzione. I pallini che attraversano il cotone procedono quasi indisturbati, ma quelli che urtano contro il diamante rimbalzano via e alcuni tornano indietro. Osservando quanti pallini diffondono nelle varie direzioni, l'ispettore può stabilire, basandosi sulle regole della geometria, la forma e le dimensioni esatte del diamante. Nello studio della struttura del protone all'AGS, i ricercatori hanno impiegato i protoni stessi come «pallini».

tali sono molto minori dell'effetto di depolarizzazione del campo di curvatura verticale sui protoni che hanno un asse di spin orizzontale. Mantenere allineati gli spin durante l'accelerazione è difficile, ma non impossibile, quando i protoni ruotano attorno a un asse verticale.

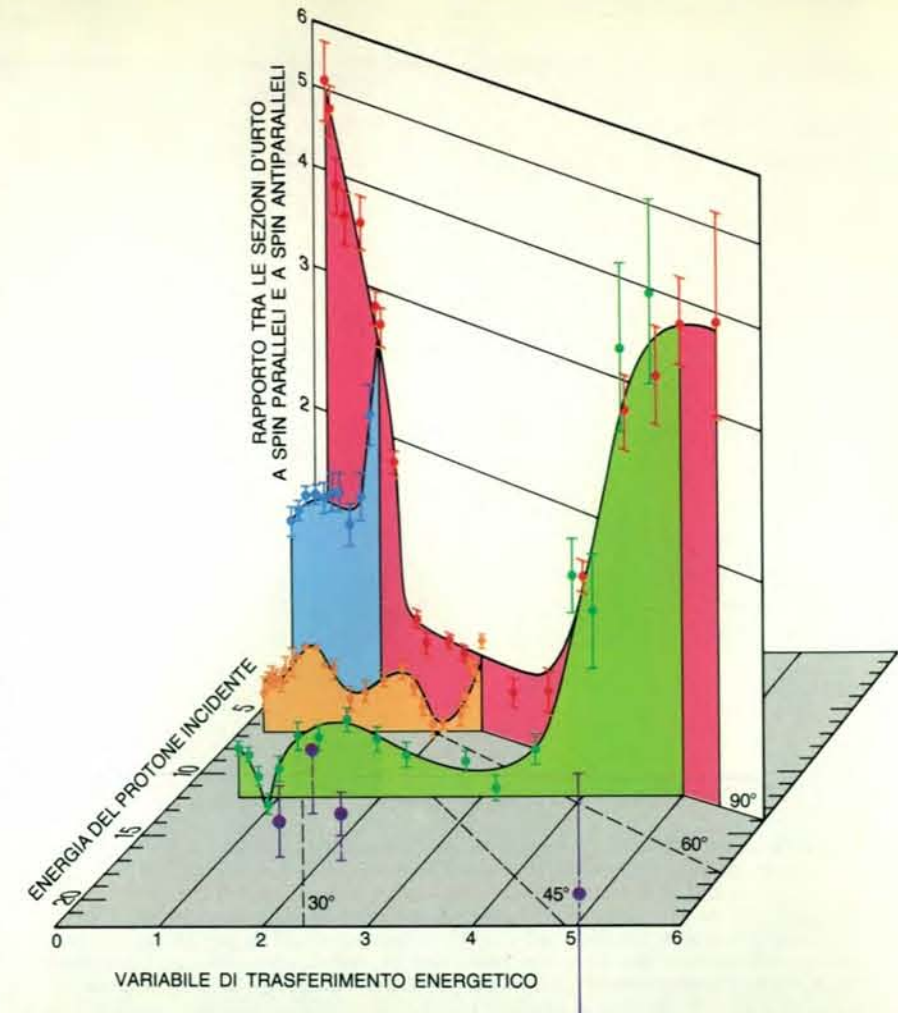
Come ho già detto, i problemi di depolarizzazione più seri sono posti dalle risonanze depolarizzanti. Quando due numeri critici, cioè il numero delle rotazioni di spin che ciascun protone compie in un giro nell'anello e il numero di volte in cui un protone oscilla in un campo magnetico orizzontale durante ogni giro, assumono lo stesso valore, viene indotta una risonanza depolarizzante. Nell'AGS, un protone subisce circa nove oscillazioni di betatrone in ogni giro nell'anello. Quando i due numeri diventano uguali, la polarizzazione può essere completamente distrutta in meno di 10 microsecondi, il tempo necessario per compiere quattro giri attorno all'AGS.

Per impedire l'insorgere di risonanze durante l'accelerazione dei protoni, abbiamo cercato di far variare uno dei due numeri nell'istante in cui essi diventavano uguali e quindi di «saltare» tutte le risonanze depolarizzanti. Abbiamo deciso di intervenire sul numero di volte in cui un protone incontra il campo magnetico orizzontale, modificando il numero di creste e di ventri nelle oscillazioni di betatrone. Era necessario, quindi, costruire particolari magneti quadrupolari impulsati, che potevano essere attivati in meno di 2,5 microsecondi. Abbiamo scoperto che magneti con una risposta così rapida non si possono realizzare in ferro comune. La scelta è caduta sulla ferrite, un materiale ceramico contenente ossido di ferro; un magnete di ferrite può generare intensi campi magnetici in tempi brevi.

I magneti furono costruiti a Michigan e i loro enormi alimentatori a Brookhaven. La costruzione dei magneti si è rivelata irta di difficoltà. La ferrite - che costa circa 130 000 lire al chilogrammo - doveva essere lavorata esattamente con profilo a iperbole, ma essa è tanto dura da poter essere tagliata solo con utensili in diamante e si frammenta se viene leggermente surriscaldata durante il taglio. Il nostro gruppo, composto di costruttori di strumenti, ingegneri, professori e studenti, cominciò a risolvere questi problemi convincendo l'US Department of Energy a fornire 1,5 tonnellate di ferrite eccedente e promuovendo l'acquisto, da parte dell'Università del Michigan, di una fresa controllata da un calcolatore. Studenti esperti di calcolatori e di geometria collaborarono con gli strumentisti che conoscevano le frese e gli utensili da taglio in diamante. Nella parte superiore della fresa, venne installata una «vasca» d'acqua in cui la ferrite era mantenuta a temperatura costante durante il taglio. La fatica si concluse con esito positivo nel 1983, quando sull'AGS furono montati 12 magneti quadrupolari impulsati. Ognuno di essi raggiunge la massima intensità del campo magnetico partendo da zero in soli 1,6 microsecondi. Per ottenere tale rapida risposta, ogni magnete richiede un enorme e complesso alimentatore che produce una corrente elettrica di 1500 ampere a 15 000 volt, con una potenza di 22,5 megawatt. Per fortuna, durante il funzionamento dell'acceleratore, il picco di potenza complessivo di tutti magneti - più di 200 megawatt - è raggiunto soltanto per pochi microsecondi ogni secondo, tanto che la potenza media risulta di circa un chilowatt. Con un lavoro intelligente e appassionato, un gruppo di ingegneri e tecnici di Brookhaven riuscì a costruire gli alimentatori. Ognuno di essi, grande circa come un autocarro, fornisce potenza a un solo magnete, che, a sua volta, ha le dimensioni di una lampada da tavolo.

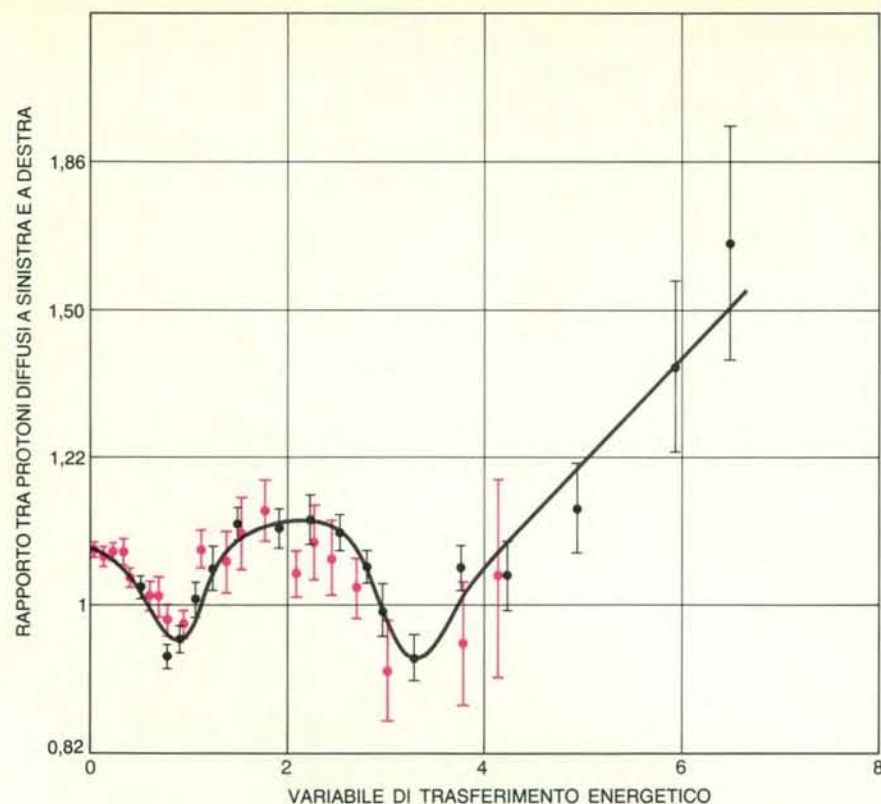
Per essere sicuri che i magneti producessero un impulso esattamente quando i protoni raggiungevano una risonanza depolarizzante, il calcolatore di controllo dell'AGS richiedeva un'accurata messa a punto. Il calcolatore registra l'esatta energia dei protoni durante ogni microsecondo del ciclo di accelerazione. Per un adeguato controllo dei protoni polarizzati, gli esperti dell'AGS dovettero migliorare la precisione strumentale di circa un fattore 10. Il miglioramento ha reso inoltre l'AGS più preciso come acceleratore di protoni non polarizzati, anche se non ancora sufficiente per individuare l'esatto istante in cui i protoni incontravano una risonanza depolarizzante. Per i ritocchi finali dovevamo affidarci allo strumento fondamentale della scienza: l'osservazione sperimentale.

Dopo tutti i nostri sforzi mentali, fisici e finanziari - descritti solo in parte - fummo un poco sorpresi e decisamente soddisfatti nello scoprire che avevamo davvero accelerato i protoni a 18,5 GeV. La polarizzazione massima era circa del 50 per cento, il che significa che il 75 per cento dei protoni ruotava in un verso e il 25 per cento nel verso opposto. L'AGS accelerava circa 20 miliardi di protoni polarizzati ogni 2,4 secondi.

A quel punto, ci occupammo ancora delle alte energie per tentare di dare una risposta definitiva alla domanda: il rapporto tra le sezioni d'urto a spin paralleli e a spin antiparalleli aumenta, diminuisce o resta fisso, per qualche misterioso motivo, esattamente al valore quattro, quando l'energia di collisione supera i 13 GeV? Purtroppo, il numero di protoni diffusi a 90 gradi diventa molto basso alle energie più elevate; la maggior parte dei protoni diffonde secondo angoli più piccoli. Per tale motivo, il nostro gruppo sperimentale, con ricercatori di Brook haven, dell'Università del Maryland, del Massachusetts Institute of Technology, della Michigan University, dell'Università di Notre Dame, del Politecnico federale di Zurigo e della Texas A&M

I valori del rapporto tra sezione d'urto a spin paralleli e a spin antiparalleli in collisioni tra protoni oscillano quando vengono modificate le caratteristiche delle collisioni. In questo grafico tridimensionale, i due assi orizzontali rappresentano l'energia dei protoni del fascio incidente e l'energia trasferita durante le collisioni. La variabile di trasferimento energetico aumenta con l'angolo di diffusione, del quale sono indicati alcuni valori. L'area in rosso rappresenta i dati di esperimenti in cui l'angolo di diffusione misurato era esattamente di 90 gradi. Le aree in blu, in arancione e in verde rappresentano esperimenti in cui l'energia dei protoni incidenti era posta a 3, 6 e 12 GeV rispettivamente. I grossi punti viola indicano dati ottenuti ad alta energia all'AGS. Le oscillazioni osservate non rientrano facilmente nel quadro delle previsioni della QCD.

University studiò per prima cosa gli angoli di diffusione più piccoli, caratterizzati da conteggi più elevati.

Ci attendeva un'altra sorpresa. A 18,5 GeV il rapporto tra le sezioni d'urto a spin paralleli e quelle a spin antiparalleli risulta molto prossimo a uno, per collisioni violente con angoli di diffusione molto minori di 90 gradi. Il grafico in cui abbiamo riportato i nostri nuovi dati indica che il rapporto può oscillare tra valori maggiori e minori di uno quando le collisioni diventano più energetiche e violente. Nicolai Tyurin dell'Istituto di Serpukhov, in Unione Sovietica, e altri teorici che avevano previsto tali oscillazioni, saranno certamente soddisfatti di questa probabile conferma delle loro ipotesi.

Tuttavia, l'andamento osservato può rivelarci qualcosa di più. In particolare, per particelle identiche, quali due protoni, l'angolo di diffusione di 90 gradi potrebbe essere un valore particolare, in corrispondenza del quale si osservano effetti singolari. L'importanza del valore di 90 gradi in questi esperimenti sullo spin fu proposto per la prima volta nel 1978 da Hans A. Bethe della Cornell University e da Victor F. Weisskopf del MIT. I nostri nuovi risultati su collisioni violente con angoli molto minori di 90 gradi si possono interpretare come una conferma dell'ipotesi di Bethe e Weisskopf. Anche se il rapporto tra le sezioni d'urto a spin paralleli e quelle a spin antiparalleli varia sensibilmente con l'energia dei protoni incidenti, per ogni energia il rapporto massimo si ha sempre a 90 gradi. Harry Lipkin del Weizmann Institute of Science, in Israele, e altri hanno proposto di recente modelli specifici, nei quali la vicinanza all'angolo di 90 gradi può fortemente aumentare il valore del rapporto. Futuri esperimenti ad

L'asimmetria sinistra-destra in eventi di diffusione a un'energia ancora più elevata, 28 GeV, rappresenta un ulteriore problema per la QCD. In questo esperimento, un fascio di protoni non polarizzati è stato diffuso da un bersaglio polarizzato. Il numero di protoni diffusi verso sinistra superava di circa due terzi il numero dei protoni diffusi verso destra. Secondo la teoria, il numero di protoni diffusi nelle due direzioni dovrebbe essere uguale. (Tale previsione sarebbe stata corretta se il bersaglio non fosse stato polarizzato.) I punti in colore indicano i dati ottenuti presso il CERN, l'Organizzazione europea per la ricerca nucleare, a un'energia di 24 GeV.

angoli di diffusione di 90 gradi con l'AGS potranno determinare la teoria corretta.

Il mistero, quindi, si infittiva a mano a mano che venivano raggiunte energie sempre più elevate. È stata forse verificata la previsione della QCD che gli effetti dello spin scompaiono alle alte energie? Un altro nuovo risultato, ottenuto a 28 GeV, mi fa ritenere che non sia così. Tale risultato è stato ricavato mentre stavamo preparando la nostra apparecchiatura per l'esperimento con il fascio polarizzato. Abbiamo calibrato gli strumenti osservando la diffusione elastica del fascio di protoni normale, non polarizzato, dell'AGS da parte del nostro bersaglio polarizzato di protoni. Secondo la previsione teorica, confermata a piccoli angoli di diffusione, il numero di protoni che diffondono a sinistra dovrebbe essere uguale al numero di protoni che diffondono a destra. Ricordo di aver pensato che, anche se la QCD non spiega i nostri primi risultati ottenuti allo ZGS, deve essere certamente corretta nel caso della semplice diffusione sinistra-destra di un fascio non polarizzato da parte di un bersaglio polarizzato.

Quando abbiamo eseguito l'esperimento, alla fine del 1983, recenti miglioramenti del nostro bersaglio di protoni polarizzati ci hanno permesso di utilizzare un fascio di quasi 100 miliardi di protoni in ogni ciclo, mantenendo ancora allineati gli spin dei protoni del bersaglio. C'erano circa 100 volte più protoni rispetto ai precedenti esperimenti di diffusione su un bersaglio polarizzato e potevamo perciò studiare le collisioni più rare e più violente. A questo punto cominciarono le sorprese. In collisioni violente a grandi angoli a 28 GeV, i protoni che diffondevano verso sinistra erano di due terzi più numerosi.

Inizialmente, osservammo un numero modesto di casi. Essendo terminato il tempo concesso per il nostro esperimento, chiedemmo una dilazione. Il comitato consultivo del programma AGS votò contro la nostra richiesta; forse i teorici del comitato credevano che avessimo osservato solo un'anomalia statistica e che ulteriori dati non avrebbero che confermato le previsioni della QCD, secondo le quali il numero di protoni diffusi verso sinistra e verso destra deve essere uguale. Il vice direttore di Brookhaven, Robert B. Palmer, prese l'insolita decisione di scavalcare il comitato. Quando, in seguito, confermammo che la diffusione mostra una dipendenza direzionale Palmer ne fu probabilmente compiaciuto.

Qual è il significato della differenza osservata tra la diffusione verso sinistra e verso destra? Forse, come suggerisce qualche teorico, la violenza e l'energia (28 GeV) delle nostre collisioni sono troppo basse perché esse possano ricadere nell'ambito di una teoria fondamentale quale la QCD. Tuttavia, potranno essere esplorate ben presto energie più alte. La differenza tra le diffusioni verso sinistra e verso destra potrà essere misurata, fra breve tempo, con i protosincrotroni da 70 a 800 GeV di Serpukhov, del CERN e del Fermilab. Il supercollisore superconduttore, o SSC, da 20 TeV (20 000 GeV), ora in fase di progetto, potrebbe consentire un ulteriore, enorme aumento di energia (si veda l'articolo *Il supercollisore a magneti superconduttori* di J. David Jackson, Maury Tigner e Stanley Wojcicki in «Le Scienze» n. 213, maggio 1986). Sembra difficile, però, aumentare di molto la violenza di collisioni elastiche: più esse sono violente, più diventano rare. Pur essendo paziente, non ho certo l'intenzione di intraprendere un esperimento che ha la probabilità di rivelare solo un evento ogni 100 anni. I miei studenti, che hanno bisogno di tali eventi per potersi laureare, sono ancora più riluttanti.

Sembra improbabile che sia possibile studiare in questo secolo collisioni elastiche ai livelli enormi di energia e di trasferimento energetico necessari per verificare adeguatamente la QCD. Forse dovremmo cercare, nel contempo, una nuova e più utile teoria delle interazioni dei protoni e dei loro costituenti rotanti. Abbiamo una conoscenza teorica migliore delle forze elettrica, debole e gravitazionale che non della forza forte (nucleare) che lega quark, protoni e neutroni. Forse la ragione di tutto ciò è che la forza nucleare è tanto intensa da rendere inutile l'approssimazione teorica chiamata teoria delle perturbazioni. Nello stesso tempo, è la stessa intensità della forza forte a renderla così importante.

Più esamino i nostri dati, più sento che essi contengono un semplice messaggio sui costituenti dei protoni, che non siamo ancora riusciti a decifrare. Non ipotizzo ciò che potrebbe accadere in futuro, dal momento che, ogni volta che esperimenti sullo spin hanno indagato regioni in precedenza inesplorate, si sono verificate sorprese, quali per esempio, l'elevata polarizzazione delle cosiddette particelle lambda prodotte in violente collisioni protone-protone e gli enormi effetti di spin osservati in collisioni protone-protone a bassa energia. Essendo io uno scienziato sperimentale, posso soltanto affidarmi al giudice definitivo della verità scientifica, l'osservazione sperimentale. Forse le misurazioni che verranno eseguite nei prossimi anni forniranno la chiave che consentirà a un abile teorico di descrivere accuratamente i costituenti del protone e le forze forti che li tengono legati.

Le cause della sindrome di Down

I geni ritenuti responsabili dell'insorgenza di molte patologie associate a questa malattia, la più comune fra quelle dovute ad anomalie genetiche, sono in corso di identificazione e localizzazione su siti del cromosoma 21

di David Patterson

La sindrome di Down è espressione di un'anomalia genetica che può avere conseguenze gravissime. La malattia è la più comune causa di ritardo mentale negli Stati Uniti, con frequenza di un caso su 700 nati vivi; i bambini colpiti da questa malattia congenita sono affetti da numerosi disturbi fisici e mentali, molti dei quali gravi. Si ritiene inoltre che i geni responsabili di alcune delle manifestazioni della sindrome di Down siano gli stessi che, in caso di espressione alterata o mal regolata dall'organismo, possono indurre svariate patologie, tra cui la leucemia e la malattia di Alzheimer, in individui altrimenti sani. Per tutte queste ragioni, la sindrome di Down è attualmente oggetto di attive ricerche che promettono di fornire una quantità di informazioni sull'espressione genetica e sulle basi molecolari della malattia.

La sindrome di Down non può essere certo considerata una malattia di recente comparsa. La sua antichità è dimostrata da un cranio sassone del IX secolo che ha le stesse dimensioni del cranio di un tipico paziente moderno affetto da sindrome di Down e da una serie di ritratti del XV secolo che raffigurano bambini con i tratti del viso caratteristici di questa malattia.

Tuttavia la sindrome non fu riconosciuta formalmente fino al 1866. In quell'anno, John Langdon Down, medico presso l'Earlswood Asylum nel Surrey, in Inghilterra, pubblicò la prima descrizione dettagliata della malattia. Il suo resoconto era basato sull'osservazione di alcuni pazienti affetti da ritardo mentale i quali presentavano un insieme caratteristico di segni clinici: in particolare, pieghe epicantiche delle palpebre, tratti del viso schiacciati, solchi palmari insoliti, ipotonia muscolare e bassa statura.

Oggi si sa che gli individui affetti dalla sindrome di Down presentano un gran numero di anomalie, sia anatomiche, sia biochimiche. Il 40 per cento di essi mostra difetti cardiaci congeniti, la maggior parte ha una massa cerebrale ridotta e molti hanno una maggior probabilità di sviluppare cataratte o altri disturbi visivi a causa di difetti del cristallino. Da un punto di vista biochimico presentano elevati livelli di purine (due delle basi azotate che formano DNA e RNA), una situazione che da sola può portare ad anomalie neurologiche, ritardo mentale e difetti del sistema immunitario. Fra le ulteriori complicanze si possono citare una maggiore predisposizione a contrarre infezioni e un aumento variabile dal 20 al 50 per cento del rischio di sviluppare leucemia.

Non stupisce, perciò, che gli individui affetti da sindrome di Down abbiano di solito una vita breve. Nel 1929 si stimava una sopravvivenza massima di soli nove anni. Nel 1980, grazie al migliore trattamento clinico, la vita media è salita a oltre 30 anni e oggi il 25 per cento degli individui affetti da sindrome di Down vive fino all'età di 50 anni.

Con l'aumento della vita media dei pazienti affetti da sindrome di Down, è emerso un altro aspetto della malattia. Negli ultimi decenni, è risultato chiaro dallo studio del materiale autoptico che tutti gli individui portatori della sindrome di Down al di sopra dei 35 anni di età sviluppano lo stesso tipo di anomalie cerebrali rilevabili microscopicamente - placche senili e ammassi neurofibrillari - dei pazienti deceduti per la malattia di Alzheimer, la più comune forma di demenza presenile. Sembra inoltre che gli individui affetti da sindrome di Down abbiano una maggiore probabilità di sviluppare la sintomatologia della malattia di Alzheimer che coinvolge le funzioni intellettive.

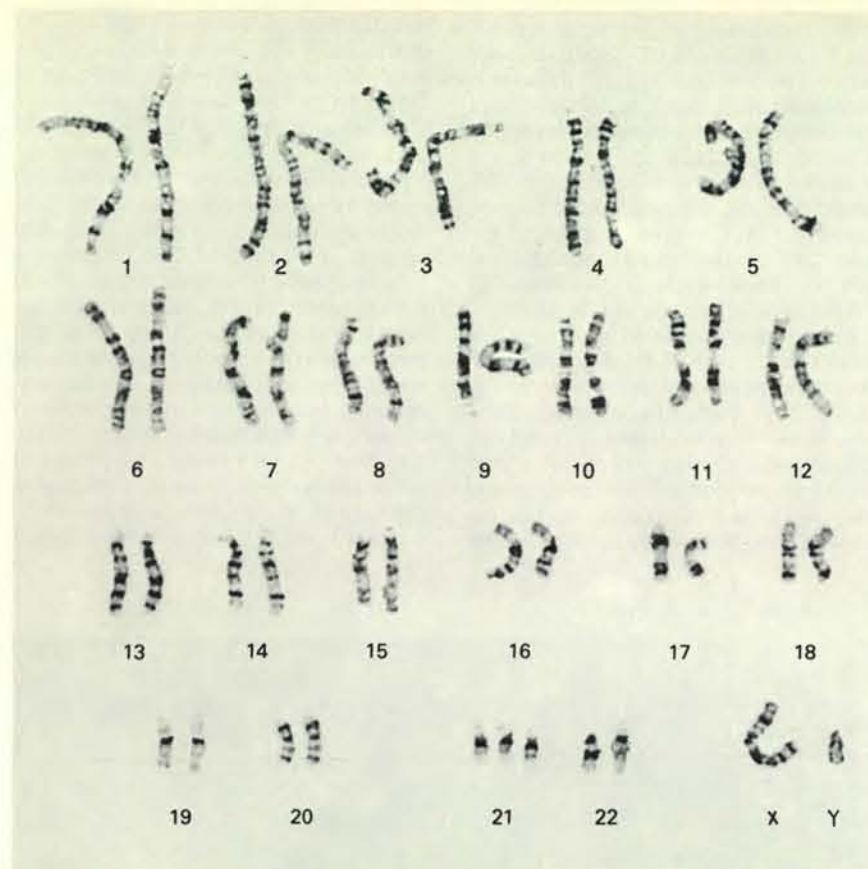
Per anni, la sindrome di Down fu una malattia di origine sconosciuta, con insorgenza apparentemente casuale. Furono proposte molte teorie, tra cui alcune che collegavano la nascita di bambini affetti da sindrome di Down alla presenza nei genitori di disfunzioni delle ghiandole endocrine, di tubercolosi o sifilide. Nel 1909, G. E. Shuttleworth del Royal Albert Asylum di Lancaster, in Inghilterra, avanzò l'ipotesi che la malattia fosse il risultato di un «esaurimento uterino». Egli basò la sua teoria sulla constatazione che molti bambini affetti da sindrome di Down sono gli ultimi nati di famiglie numerose. Shuttleworth aveva parzialmente ragione: i bambini portatori della sindrome di Down sono spesso gli ultimi di una numerosa progenie, ma il fatto viene oggi attribuito all'età avanzata della madre piuttosto che al numero dei figli concepiti.

All'inizio degli anni trenta, sia Adrian Bleyer, della School of Medicine della Washington University di St. Louis, sia P. J. Waardenburg suggerirono indipendentemente che la sindrome di Down potesse essere associata alla non disgiunzione, cioè all'errata ripartizione dei cromosomi durante la meiosi (il processo di divisione cellulare che porta alla produzione delle cellule uovo e degli spermatozoi). La loro intuizione derivava dall'osservazione del comportamento dei cromosomi di una pianta del genere enotera, la cui errata separazione portava a una progenie sterile con 15 cromosomi invece dei normali 14. Se la non disgiunzione poteva condurre ad anomalie nell'enotera, non poteva essere ugualmente responsabile delle anomalie associate alla sindrome di Down? Gli sforzi per provare questa ipotesi furono ostacolati dal fatto che nessuno aveva ancora effettuato un conteggio corretto dei cromosomi umani.

Solo negli anni cinquanta Joe Hin Tjio e Albert Levan dell'Istituto di genetica di Lund, in Svezia, furono in grado di stabilire in 46 il numero esatto dei cromosomi umani, confermando il legame tra la sindrome di Down e la non disgiunzione. Jérôme Lejeune, Marthe Gautier e Raymond Turpin dell'Institut de Progénèse di Parigi contarono i cromosomi in cellule di pazienti affetti da sindrome di Down. Essi osservarono che in questi pazienti si era verificata una non disgiunzione dei cromosomi: tutti presentavano nelle loro cellule tre copie del cromosoma 21 anziché due, per un totale di 47 cromosomi invece di 46.

La non disgiunzione dei cromosomi è dovuta a una disfunzione della divisione cellulare. Qualunque cromosoma essa coinvolga, le sue conseguenze sono spesso gravi. Nella non disgiunzione, la coppia di cromosomi che normalmente si separa nelle due cellule figlie durante la meiosi non si divide correttamente, così che ciascuna cellula figlia riceve o entrambi i cromosomi o nessuno. Le cellule che mancano di un cromosoma di solito muoiono; quelle che hanno due cromosomi ne ricevono un terzo durante la fecondazione, risultando trisomiche. I feti che presentano trisomia raramente sopravvivono e quelli che vi riescono sono affetti da numerose anomalie biochimiche e fisiche. La sindrome di Down, o trisomia 21, è di gran lunga la più comune trisomia dell'uomo osservabile nei nati vivi, forse perché il cromosoma 21 è il più piccolo cromosoma umano.

La trisomia di almeno una parte del cromosoma 21 si è finora rivelata il fattore determinante della sindrome di Down. Non vi sono individui con i sintomi clinici della sindrome di Down che non abbiano almeno una trisomia parziale del cromosoma 21 così come non vi sono casi conosciuti di individui con trisomia 21 che non siano affetti da sindrome di Down. L'anomalia non sempre deriva dalla non disgiunzione; il 5 per cento dei pazienti affetti da sindrome di Down ha un tipo diverso di mutazione cromosomica chiamato traslocazione, in cui soltanto una parte del cromosoma 21 è presente in tre copie. Questo accade quando una parte del cromosoma 21 si unisce a un altro cromosoma, in genere i cromosomi 13, 14, 15, 21 o 22. Quando accade questo, si verificano difficoltà di appaiamento durante la meiosi e il frammento di cromosoma 21 appare in una delle cellule figlie insieme a un cromosoma 21 normale; come nella non disgiunzione, il frammento diventa trisomico durante le fecondazione. Indagini cliniche svolte su un piccolo gruppo di pazienti affetti da sindrome di Down come risultato di una traslocazione hanno dimostrato che la trisomia del solo terzo

Questa bambina di quattro anni, affetta da sindrome di Down in forma attenuata, mostra i tratti fisici caratteristici della malattia, tra cui pieghe epicantiche delle palpebre, volto allargato e naso schiacciato. La bambina è stata molto avvantaggiata dall'avere una famiglia che si è sempre attivamente occupata di lei. Già all'età di quattro settimane aveva cominciato una terapia regolare; a poco più di un anno ha iniziato la cura dei difetti del linguaggio. I risultati hanno avuto una positiva influenza sul suo sviluppo intellettuale e fisico. Alcuni aspetti tipici della sindrome di Down, tra cui bocca aperta, lingua sporgente e scarsa coordinazione, che sono associati a un ridotto tono muscolare, sono molto meno evidenti nel suo caso, come conseguenza della continua terapia educativa e fisica. Molti individui affetti da sindrome di Down, che si trovano in una situazione analoga, hanno una speranza di vita molto maggiore e sono capaci di raggiungere un notevole grado di indipendenza.

Il cariotipo è il corredo cromosomico completo di un organismo. I singoli cromosomi visibili nell'illustrazione sono stati ritagliati da una microfotografia ottenuta durante la metafase. I cromosomi sono appaiati e disposti in coppie omologhe. Questa tecnica consente di compiere un conteggio accurato dei cromosomi di una cellula e di esaminarli alla ricerca dei difetti più grossolani. Nell'uomo vi sono normalmente 46 cromosomi (23 coppie appaiate), benché questo numero possa variare a causa della non disgiunzione. Le tre copie del cromosoma 21 visibili indicano che la persona da cui sono state ottenute le immagini cromosomiche è affetta da sindrome di Down.

inferiore del cromosoma 21 è sufficiente a causare la malattia.

Il cromosoma 21 contiene soltanto 45 milioni di coppie di basi del DNA (su un totale di 3 miliardi di coppie di basi nel nucleo di una cellula umana), ovvero l'1,5 per cento del totale del materiale genetico. Si calcola che gli esseri umani abbiano circa 100 000 geni funzionali. Assumendo che ogni segmento della molecola di DNA contenga circa lo stesso numero di geni, si può calcolare che il cromosoma 21 abbia circa 1500 geni. Meno di 20 di essi sono stati identificati.

I cromosomi sono costituiti da due parti identiche chiamate cromatidi, che si congiungono nel centromero. Il braccio corto del cromosoma che sta al di sopra del centromero è conosciuto come regione p; il braccio più lungo al di sotto del centromero è noto come regione q. Si possono visualizzare bande alternate chiare e scure trattando i cromosomi con particolari coloranti. Le bande vengono identificate a seconda che siano sui bracci p o q del cromosoma e vengono numerate in relazione alla loro distanza dal centromero; più distante è la banda, più alto è il numero. Le bande hanno una larghezza variabile e ciascun cromosoma mostra una distribuzione caratteristica. Il gene che codifica per gli enzimi necessari alla sintesi delle purine, per esempio, si trova sulla banda q22.1 del cromosoma 21; questo significa che il gene è localizzato in una banda della metà inferiore del braccio q. Le bande più piccole di un cromosoma umano che possono essere distinte al microscopio comprendono da 2 a 5 milioni di coppie di basi del DNA e perciò possono contenere molti geni.

Nell'affrontare lo studio del cromosoma 21, i ricercatori hanno tentato di rispondere a quattro domande. Quali geni occupano la regione del cromosoma 21 direttamente responsabile della sindrome di Down? Quali di essi sono responsabili della patogenesi della malattia? Quali proteine sono codificate dai geni responsabili? Mediante quale meccanismo la presenza di tre copie dei geni (e non due) provoca la sindrome di Down?

Per rispondere a queste domande si deve provvedere alla localizzazione e all'identificazione dei diversi geni del cromosoma 21. Questo procedimento è conosciuto come mappatura genica. Le mappe geniche ad alta risoluzione localizzano i geni in regioni circoscritte del cromosoma, e precisamente in specifici frammenti di DNA, e, di conseguenza, forniscono il maggior numero di informazioni.

Per ottenere mappe a così alta risoluzione sono normalmente indispensabili diversi passaggi ed è necessario tenere conto delle informazioni ottenute con differenti tecniche citogenetiche e biochimiche. I geni possono talvolta essere mappati su un cromosoma per via indiretta, mediante indagini biochimiche basate sul dosaggio dei prodotti genici. La teoria è semplice: dal momento che gli individui che presentano una trisomia possiedono tre copie di ciascun gene invece di due, essi devono produrre un 50 per cento in più di un particolare prodotto genico rispetto agli individui normali. Perciò rilevare in un paziente una quantità di una particolare proteina o un'attività enzimatica equivalente a circa una volta e mezzo quella normale costituisce un buon indizio del fatto che la proteina o l'enzima è codificato da un gene su un cromosoma trisomico.

Il metodo è particolarmente utile nei casi di trisomia parziale, quando l'aumento dell'espressione genetica può essere associato alla presenza di una piccola porzione del cromosoma. In alcuni casi, è possibile mappare, mediante questo metodo, un gene su una singola banda del cromosoma. Tuttavia la correlazione tra un prodotto genico e la presenza di un cromosoma soprannumerario non è sempre corretta, non solo perché i livelli dei prodotti genici variano da individuo a individuo o da tessuto a tessuto, ma anche perché copie multiple di uno stesso gene non sono sempre ugualmente funzionali; una reazione autoregolata nota come compensazione del dosaggio tende a bilanciare l'aumento della produzione proteica. È perciò importante confermare la localizzazione di un gene identificato sulla base del dosaggio genico usando altri metodi, quando ciò è possibile.

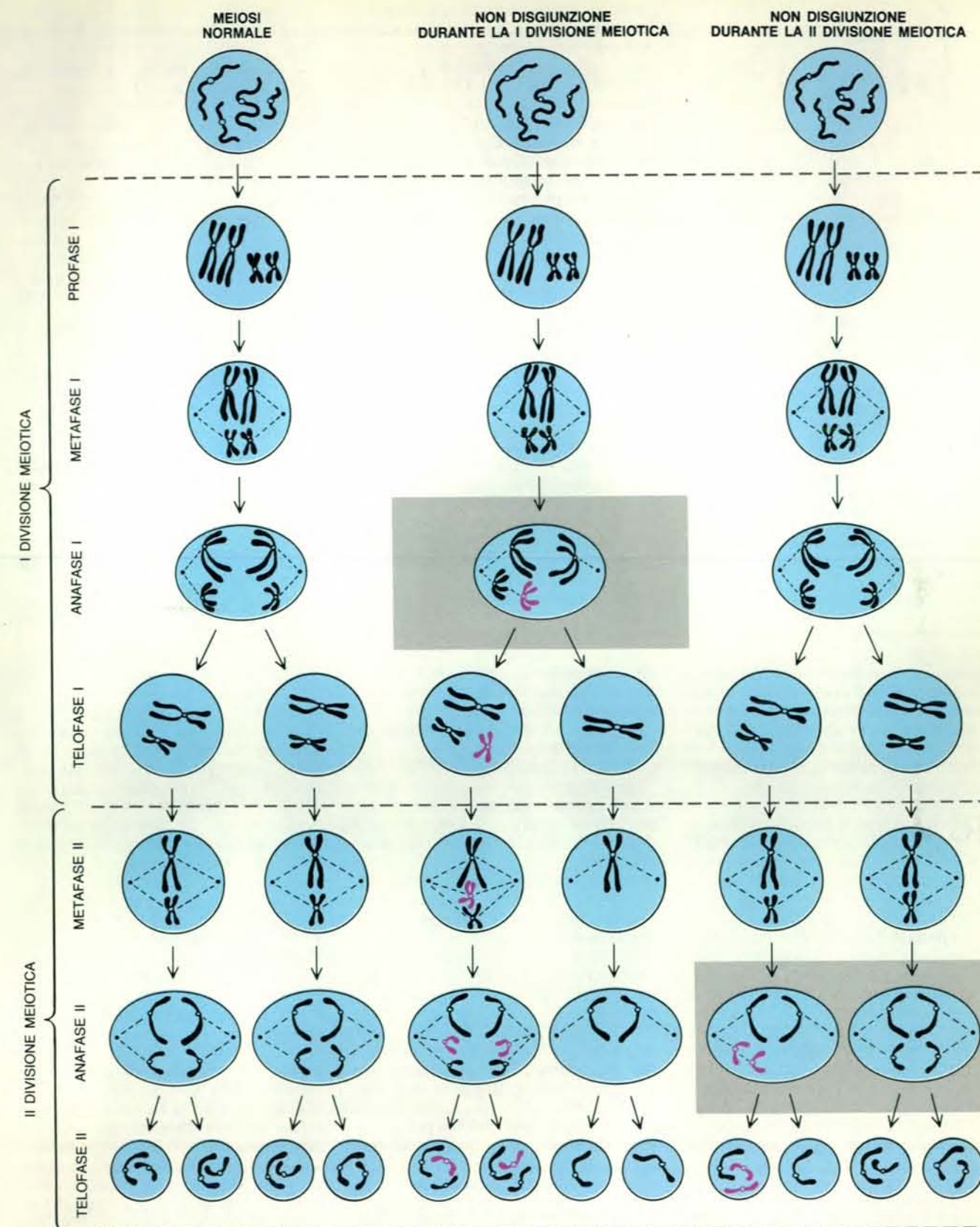
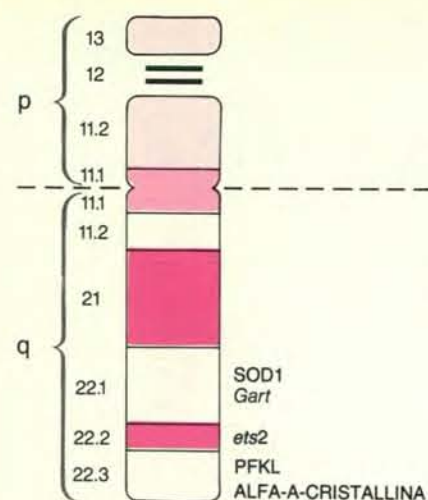
Un'altra importante tecnica di mappatura utilizza i marcatori genetici; si tratta di variazioni genetiche, espresse nella composizione proteica e nell'aspetto fisico del fenotipo, che possono essere rintracciate di generazione in generazione. Se un marcatore genetico, sicuramente codificato dal DNA del cromosoma 21, compare quasi sempre associato a un secondo tratto fenotipico, allora si può ritenere che anche questo secondo tratto si trovi sul cromosoma 21. Quando entrambi i geni sono mappati sullo stesso cromosoma è possibile stabilire la loro distanza relativa osservando quanto spesso essi si separino durante la meiosi, quando i cromosomi omologhi si scambiano frammenti nel corso della ricombinazione genetica. I geni che sono associati (che cioè, di solito, vengono trasmessi insieme) si trovano probabilmen-

MEIOSI NORMALE

NON DISGIUNZIONE DURANTE LA I DIVISIONE MEIOTICA

NON DISGIUNZIONE DURANTE LA II DIVISIONE MEIOTICA

I DIVISIONE MEIOTICA: PROFASE I, METAFASE I, ANAFASE I, TELOFASE I

II DIVISIONE MEIOTICA: METAFASE II, ANAFASE II, TELOFASE II

La meiosi normale (*a sinistra*) porta alla produzione di cellule uovo e spermatozoi aploidi, contenenti un corredo cromosomico dimezzato rispetto alle altre cellule. Durante la I divisione meiotica i cromosomi omologhi appaiati si scambiano l'informazione genetica (nella profase) e si separano agli estremi opposti della cellula (anafase), dopo di che la cellula si divide (telofase). Il processo si ripete durante la II divisione meiotica, quando i due cromatidi fratelli si separano e ogni cellula si divide di nuovo per formare quattro cellule figlie. Si ha non disgiunzione quando i cromosomi non si separano durante l'anafase o della I divisione meiotica (*al centro*) o della II (*a destra*). I cromosomi, che sono normalmente spinti dalle fibrille del fuso agli estremi opposti della cellula, rimangono insieme e due cromosomi omologhi appaiati (nel primo caso) oppure due cromatidi fratelli (nel secondo) vanno verso la stessa estremità della cellula. Si ha così un'anomalia nel numero dei cromosomi delle cellule figlie. Se la non disgiunzione interessa il cromosoma 21, gli individui nati con tre copie di esso sono affetti da sindrome di Down.

Cinque geni probabilmente associati alla sindrome di Down sono stati mappati in una regione limitata del braccio lungo, o braccio q, del cromosoma 21. I geni che codificano per le proteine superossidodismutasi (SOD1) e alfa-A-cristallina sono localizzati rispettivamente sulle bande q22.1 e q22.3, e i geni *Gart* ed *ets*2 sulle bande q22.1 e q22.2. Anche un gene che codifica per l'enzima epatico fosfofruttochinasi (PFKL) è localizzato sulla banda q22.3.

te molto vicini l'uno all'altro sul cromosoma; i geni non associati (quelli che, in genere, si muovono indipendentemente durante la ricombinazione) sono di solito abbastanza distanti fra loro sul cromosoma o si trovano addirittura su cromosomi diversi. Perciò la frequenza di ricombinazione costituisce un parametro per misurare la distanza fra geni.

Recentemente, è stato introdotto un nuovo, efficace metodo di studio chiamato mappatura dei frammenti di restrizione, o mappatura RFLP (dall'inglese *restriction fragment length polymorphism*). I frammenti di restrizione sono pezzi di DNA che sono stati tagliati da enzimi chiamati endonucleasi di restrizione, ciascuno dei quali riconosce specifiche sequenze su un filamento di DNA e coerentemente taglia il DNA solo in questi punti. Mutazioni nel DNA possono far variare la distribuzione dei siti di taglio, producendo frammenti di lunghezze diverse da quelle normalmente presenti. Questi cambiamenti di lunghezza (gli RFLP) permettono di individuare differenze in regioni riconoscibili del filamento di DNA. Gli RFLP non si verificano necessariamente in regioni che codificano per determinati geni, ma sono tuttavia marcatori estremamente efficaci ai quali possono essere associati geni o altri marcatori genetici. Questo metodo è stato usato estesamente per mappare geni e sequenze di DNA sul cromosoma 21.

La possibilità di isolare geni specifici o sequenze di DNA ha anche reso possibile una nuova tecnica nota come ibridazione *in situ*. Sequenze specifiche di DNA possono essere isolate, marcate con un radioisotopo e usate per esaminare una preparazione di cromosomi di un altro individuo o di un'altra specie. Il «frammento sonda» ricerca i filamenti di DNA nei cromosomi che sono complementari a esso e vi si lega. Con la preparazione dei cromosomi si può allora impressionare un'emulsione fotografica per determinare con precisione dove i radioisotopi si siano legati ai cromosomi intatti. Questa tecnica permette una risoluzione abbastanza elevata: frammenti di DNA possono essere localizzati con una precisione pari alla lunghezza di un terzo del cromosoma 21.

Il campo della genetica molecolare è stato rivoluzionato negli ultimi anni dallo sviluppo di una tecnica di mappatura estremamente efficace nota come ibridazione interspecifica delle cellule somatiche. I genetisti hanno sviluppato metodi grazie ai quali cellule umane possono essere fuse con cellule di roditori (generalmente cellule di topo o di criceto) per creare ibridi contenenti cromosomi di entrambe le specie. Per ragioni non ancora chiare, la maggior parte dei cromosomi umani viene eliminata da queste cellule ibride, mentre i cromosomi del roditore sono conservati. È possibile, perciò, ottenere una cellula ibrida in cui un particolare cromosoma sia l'unico cromosoma umano presente. Poiché il cromosoma umano dell'ibrido è almeno in parte funzionale, è spesso possibile differenziare i prodotti proteici da esso codificati da quelli codificati dai cromosomi del roditore.

Francis H. Ruddle e collaboratori alla Yale University furono i primi ricercatori a sviluppare un ibrido di questo genere contenente soltanto il cromosoma 21 delle cellule umane e l'assetto cromosomico completo di una linea selezionata di cellule di topo, chiamate cellule A9. Da allora molti laboratori hanno ottenuto ibridi contenenti il cromosoma 21 mediante la stessa metodica.

All'Eleanor Roosevelt Institute for Cancer Research di Denver, abbiamo utilizzato una linea cellulare ricavata da ovaio di criceto (CHO) invece di cellule A9 di topo. Queste cellule CHO sono ideali per studi di ibridazione cellulare del cromosoma 21 poiché portano una mutazione che inattiva il gene *Gart*, che, negli esseri umani, è localizzato sul cromosoma 21. Il gene codifica per un enzima chiamato fosforibosilglicinammidesintetasi. Se manca questo enzima, le cellule CHO devono essere coltivate in un terreno contenente purine, senza le quali morirebbero. Se si ibridano cellule CHO e cellule umane, ciascuna cellula ibrida contenente il cromosoma 21 sarà in grado di crescere in un terreno privo di purine poiché avrà acquisito il gene *Gart* umano. In questo modo possiamo selezionare gli ibridi cellulari contenenti il cromosoma umano 21.

Alcune varianti di questa tecnica hanno consentito di individuare il gene *Gart* in un locus del cromosoma 21, designato 21q22, creando ibridi tra cellule CHO e cellule umane contenenti una traslocazione del cromosoma 21. Attraverso l'analisi citogenetica dei cromosomi portatori della traslocazione nelle cellule ibride, siamo stati in grado di suddividere il cromosoma 21 in alcuni frammenti e di assegnare marcatori genetici a un frammento contenente il gene *Gart*. L'unico aspetto negativo di questo procedimento di mappatura è il fatto che esso dipende da traslocazioni che avvengono spontaneamente e che coinvolgono diversi frammenti del cromosoma 21, traslocazioni che, in effetti, sono assai rare.

Un metodo per ovviare alla scarsità di frammenti di traslocazione è di spezzare il cromosoma 21 esponendolo a forti dosi di radiazioni. I frammenti ottenuti possono essere usati per formare ibridi con cellule di criceto e manipolati come i frammenti naturali. Gli ibridi possono essere analizzati cercando specifici geni o sequenze di DNA, e si possono formare gruppi di associazione molecolare con il gene *Gart* come elemento centrale.

Il passo finale nella creazione di una mappa ad alta risoluzione del cromosoma 21 è la misurazione della distanza fisica tra i vari geni e i marcatori, mediante il conteggio del numero di coppie di basi presenti fra essi. Lo strumento fondamentale è la tecnica dell'elettroforesi, con la quale si espongono frammenti di DNA a un campo elettrico. Poiché il DNA ha una carica negativa proporzionale alle sue dimensioni, le molecole di DNA si muovono verso l'elettrodo positivo. I frammenti più piccoli si muovono abbastanza rapidamente, quelli grandi più lentamente. Le dimensioni dei frammenti, misurate in numero di coppie di basi, sono indicate dalla posizione finale dei frammenti sul substrato di gel.

David C. Schwartz e Charles R. Cantor del Columbia University College of Physicians and Surgeons sono riusciti a modificare la tecnica standard dell'elettroforesi analizzando il DNA con campi elettrici pulsanti in direzione diagonale. Per ragioni non ancora chiaramente comprese, la tecnica permette di muovere e di separare l'uno dall'altro frammenti di DNA molto grandi (di grandezza variabile da 50 000 a più di cinque milioni di coppie di basi). La normale elettroforesi consente l'analisi di frammenti non più grandi di 50 000 coppie di basi. Tuttavia i geni hanno dimensioni molto variabili. Molti sono assai più grandi di 50 000 coppie di basi; si calcola che alcuni raggiungano dimensioni di un milione di coppie di basi o più. La tecnica del campo pulsante consente ai genetisti di analizzare l'intero gene invece che un solo frammento e anche grossi pezzi di DNA che possono contenere molti geni.

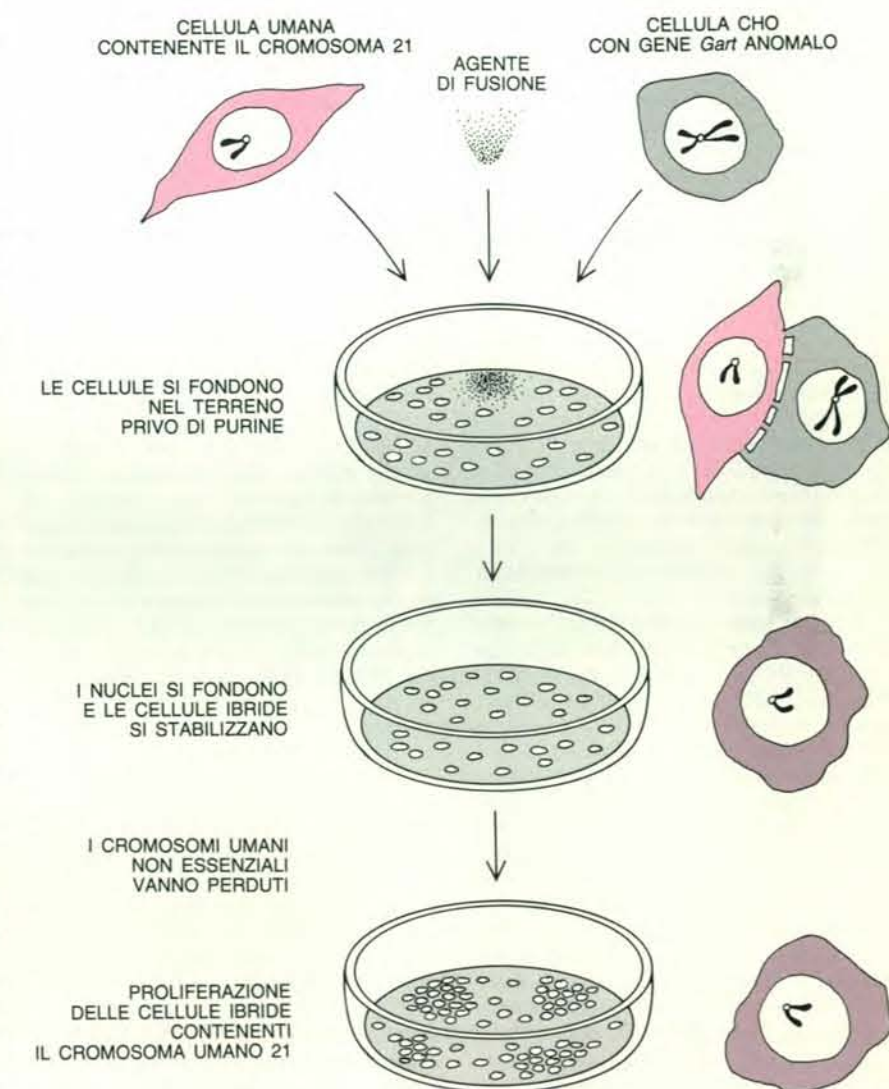
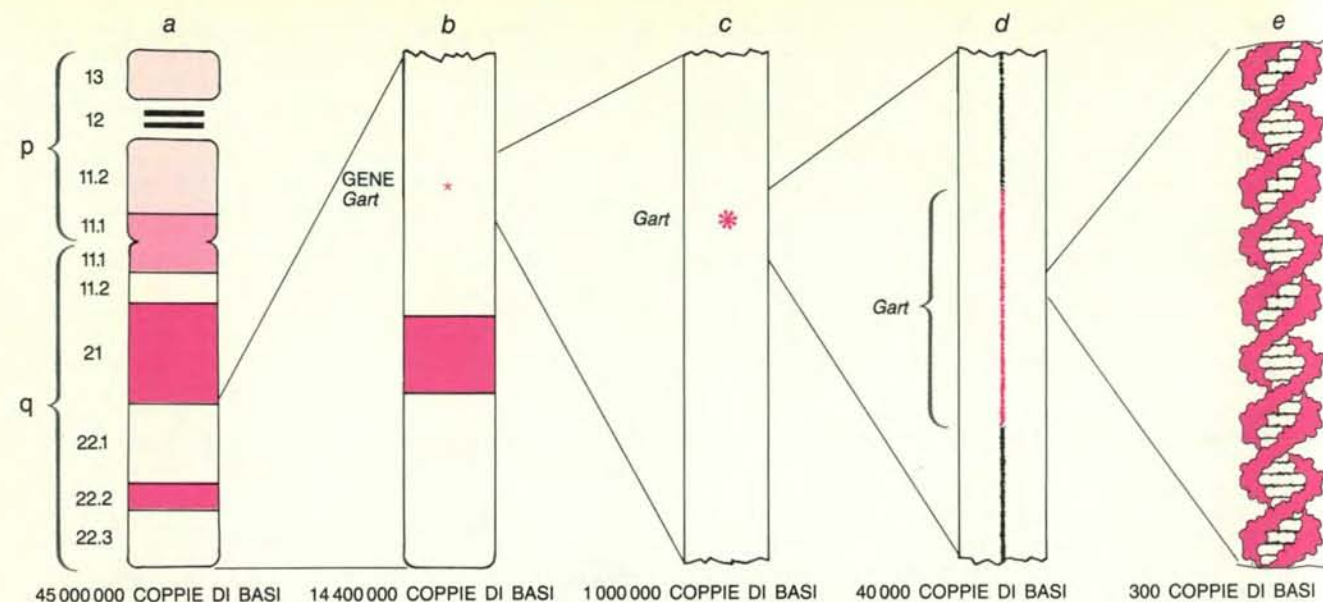
Il confronto fra le mappe ottenute con questi metodi rivela una generale coerenza tra di esse, almeno per quanto riguarda l'ordine dei frammenti di DNA e dei marcatori sul cromosoma 21. Sommando le informazioni ottenute con tali metodi e con vari studi clinici e citogenetici, i ricercatori hanno identificato una quantità di geni localizzati sulla banda q22 del cromosoma 21. Sembra che questa sia la regione del cromosóma direttamente responsabile della patogenesi della sindrome di Down.

Poiché lo scopo di queste ricerche è quello di correlare determinati geni a patologie specifiche e dal momento che i geni codificano per le proteine (che, in ultima analisi, determinano il fenotipo dell'essere umano), è chiaramente importante sapere quali proteine contribuiscano alla patologia della sindrome di Down. Una volta identificate le proteine e riconosciuti i loro effetti, è presumibile che in futuro possano essere prese misure per contrastare l'azione deleteria di queste proteine.

Esempi di geni mappati sulla banda q22 del cromosoma 21 sono il gene *Gart* sopra menzionato, il gene *ets*2 e i geni che codificano per le proteine alfa-A-cristallina, superossidodismutasi e beta amiloide. Questi geni consentono di intravvedere la complessa relazione tra di essi (e le proteine da essi codificate) e lo sviluppo e la malattia nell'uomo. Per esempio, recenti sviluppi nel campo della biologia molecolare hanno gettato luce sulla possibile relazione tra il gene *ets*2 sul cromosoma 21 e l'instaurarsi della leucemia.

Da anni i medici sapevano che gli individui affetti da sindrome di Down avevano una elevata probabilità di contrarre la leucemia, ma avevano difficoltà a spiegare i meccanismi per cui ciò avveniva. Gli oncologi scoprirono che il 18 per cento degli individui affetti da un particolare tipo di leucemia, la leucemia acuta mieloblastica, sottogruppo M2, ha una traslocazione cromosomica reciproca che coinvolge frammenti dei cromosomi 8 e 21. Essi non sono affetti da sindrome di Down; solo le loro cellule leucemiche sono interessate dalla traslocazione, mentre i cromosomi delle altre cellule sembrano essere normali. Grazie ai metodi di mappatura citogenetica, i ricercatori hanno localizzato il punto di rottura sul cromosoma 21 nella regione q22. Un'altra scoperta affascinante è che la trisomia 21 è l'anomalia cromosomica osservata più frequentemente nelle cellule leucemiche, specie nei bambini. Nel 1985 Janet D. Rowley e colleghi della Pritzker School of Medicine dell'Università di Chicago, in collaborazione con il nostro gruppo, ipotizzarono che la leucemia potesse essere causata dall'alterazione dell'attività di uno o più oncogèni sul cromosoma 21. Il nostro gruppo ha proposto che la responsabilità dell'alterazione potesse essere attribuita o alla trisomia del cromosoma 21 o a una traslocazione di materiale genetico dal cromosoma 21 al cromosoma 8.

Questa ipotesi sta diventando via via sempre più credibile. Takis S. Papas del National Cancer Institute, in collaborazione con molti altri ricercatori, compresi alcuni membri del nostro gruppo, ha dimostrato che in effetti esiste un oncogène sul cromosoma 21, chiamato oncogène *ets*2, e lo ha localizzato nella regione interessata dalla traslocazione 8; 21. Sono attualmente in corso studi per confrontare l'espressione del gene *ets*2 in cellule normali con l'espressione del gene in cellule divenute trisomiche o in cui sia avvenuta una traslocazione.

La scoperta che il gene *Gart*, che codifica per tre diversi enzimi coinvolti nella sintesi delle purine, si trova nella regione q22 del cromosoma 21 può spiegare perché individui affetti da sindrome di Down abbiano nel siero livelli di purine superiori al normale. Dal momento che elevati livelli di purine sono associati a una vasta gamma di patologie, tra cui il ritardo mentale, si può ipotizzare che la trisomia di questo singolo gene sul cromosoma 21 sia sufficiente per spiegare molte delle manifestazioni associate alla sindrome di Down, tra cui, appunto, il ritardo mentale. I ricercatori sperano che il ritorno a normali livelli di purine attraverso il controllo dell'espressione del gene *Gart* (forse nei primi stadi dello sviluppo fetale) possa un giorno condurre a un miglioramento della sintomatologia della sindrome di Down.

L'ibridazione delle cellule somatiche tra cellule umane e cellule d'ovaio di criceto (CHO) può portare alla creazione di ibridi che contengono cromosomi delle due specie. Un agente che favorisce la fusione cellulare, come il virus Sendai o il polietilenglicole, è aggiunto al terreno di coltura per promuovere la fusione. Le cellule umane possono essere ibridate con cellule CHO che non possono sintetizzare purine per una anomalia del gene *Gart*. I cromosomi umani non essenziali alla crescita dell'ibrido sono gradualmente eliminati durante la divisione cellulare. Se l'ibridazione avviene su un terreno privo di purine, solo le cellule che contengono il gene umano *Gart* (sul cromosoma 21) e possono sintetizzare purine sopravvivono. Si possono così identificare le cellule ibride che contengono tutto il cromosoma 21 o una parte di esso. Se solo una parte del cromosoma è conservata, si può localizzare la regione del cromosoma che contiene il gene *Gart*.

L'analisi molecolare del cromosoma 21 viene effettuata a differenti livelli di risoluzione, permettendo di mappare un gene come il *Gart* in una regione del cromosoma e anche di determinare la sequenza di coppie di basi. L'analisi citogenetica si basa su uno o due metodi: regioni di DNA di lunghezza variabile da uno a due milioni di coppie di basi sono identificate trattando i cromosomi con coloranti (*a*) oppure si può ottenere un più alto livello di risoluzione se è avvenuta una traslocazione che ha coinvolto soltanto un frammento del cromosoma 21 (*b*). Un livello di risoluzione ancora più elevato si può avere con l'utilizzo di tecniche più sofisticate, tra cui la creazione di vari ibridi interspecifici tramite l'irradiazione e l'uso dell'elettroforesi a campo pulsante (*c*), che spingono il livello di risoluzione da 10 milioni a 30 000 coppie di basi. Tecniche di clonazione molecolare e la normale elettroforesi su gel permettono una risoluzione che va da 50 000 fino ad alcune centinaia di coppie di basi (*d*). Infine, la sequenza di coppie di basi in un dato gene può essere determinata con tecniche di sequenziazione del DNA.

L'aumento del rischio di insorgenza di cataratta e di difetti del cristallino in individui affetti da sindrome di Down può essere anch'esso spiegato con l'espressione anormale di una particolare proteina. I ricercatori hanno scoperto che il gene che codifica per la proteina alfa-A-cristallina, un componente strutturale del cristallino, si trova sul cromosoma 21 nella regione q22.

Un altro gene identificato sul cromosoma 21 è quello che codifica per la forma solubile dell'enzima superossidodismutasi (SOD1). Questo enzima fa parte di un complesso sistema che protegge le cellule dei mammiferi dai radicali liberi contenenti ossigeno: molecole estremamente reattive, rilasciate durante l'ossidazione, che sono forse implicate nel processo di invecchiamento naturale. Alterazioni dei livelli di SOD1 possono contribuire al ritardo mentale in pazienti affetti da sindrome di Down e possono anche spiegare il loro invecchiamento precoce. L'espressione dell'enzima sembra essere correlata al dosaggio dei prodotti genici e il gene responsabile è stato localizzato nella regione q22 del cromosoma 21. La comprensione del funzionamento di questo gene potrà fornire ai ricercatori utili chiarimenti sui normali processi di invecchiamento nell'uomo.

Una recente, entusiasmante scoperta è la correlazione genetica tra la malattia di Alzheimer e la sindrome di Down. George G. Glenner e colleghi dell'Università della California a San Diego hanno dimostrato che la proteina beta amiloide (uno dei principali componenti delle placche neurofibrillari che si accumulano nel cervello degli individui affetti dalla malattia di Alzheimer) è identica alla proteina che si accumula nelle lesioni cerebrali, apparentemente uguali, presenti nel cervello degli individui affetti da sindrome di Down sopra i 35 anni di età. Ricercatori che lavorano in diversi laboratori hanno dimostrato contemporaneamente che il gene che codifica per la proteina beta amiloide si trova sul cromosoma 21. Inoltre hanno scoperto che anche il gene ritenuto responsabile di una forma ereditaria della malattia di Alzheimer è localizzato sul cromosoma 21. Non si sa ancora, però, se questi due geni siano di fatto un unico gene.

Molti progressi sono stati compiuti nell'identificazione e nella mappatura di vari geni sul cromosoma 21, un processo che certamente si intensificherà in futuro. Tuttavia, non abbiamo ancora una prova completa che la trisomia anche di una sola parte del cromosoma 21 sia direttamente responsabile delle patologie descritte in questo articolo. Si spera che i nuovi progressi dell'ingegneria genetica, che rendono possibile la manipolazione di animali «modello», possano rimediare a questa situazione.

I genetisti hanno già scoperto che una parte del cromosoma 16 del topo contiene i geni *ets*2, *Gart* e SOD1, insieme a molte altre sequenze di DNA, che, nell'uomo, sappiamo essere localizzate sul cromosoma 21. John D. Gearhart della Johns Hopkins University School of Medicine e Charles J. Epstein e David R. Cox della School of Medicine dell'Università della California a San Francisco e colleghi stanno tentando di riprodurre topi portatori di una trisomia parziale del cromosoma 16 per vedere se essi sviluppano alcune delle patologie associate alla sindrome di Down. Benché i topi riprodotti in questo modo difficilmente sviluppino le stesse patologie dell'uomo, è ragionevole pensare che i meccanismi di base dell'espressione genetica siano gli stessi, così che la ricerca con i topi fornirà ai genetisti nuovi suggerimenti sulla patologia della sindrome di Down.

Un altro strumento molto promettente per la genetica molecolare è una tecnica di ingegneria genetica che porta alla creazione di topi transgenici, che contengono cioè singoli geni o gruppi di geni di origine umana. La fattibilità di questa tecnica è stata ampiamente dimostrata per gli oncogèni, i geni del fattore della crescita e geni di enzimi metabolici; ora è tempo di applicarla ai geni del cromosoma 21 che si ritiene siano associati alla sindrome di Down.

In futuri esperimenti, geni del cromosoma 21 quali *ets*2, SOD1 e *Gart* saranno iniettati in cellule di embrioni di topo e si esprimeranno quindi nella prole. In tal modo gli effetti della trisomia di un solo gene o di un piccolo gruppo di geni potranno essere studiati in dettaglio, fornendo una quantità di informazioni sui molteplici effetti di un singolo gene. Sarà anche possibile sperimentare, in questi animali «modello», vari metodi per prevenire e curare alcune delle patologie associate alla sindrome di Down.

La teoria della selezione clonale

Sono stati necessari quasi cento anni per definire e comprovare una teoria che spiegasse la formazione degli anticorpi, una vasta gamma di proteine in grado di difendere l'organismo dall'invasione di sostanze estranee

di Gordon L. Ada e Sir Gustav Nossal

Come possono le cellule produrre una gamma così ampia di anticorpi ? Oggi a questa domanda è stato in gran parte risposto e le proteine con funzione difensiva, prodotte dal sistema immunitario, sono al servizio della medicina e della ricerca. Tuttavia, ancora non molti anni fa, gli anticorpi erano oggetti misteriosi e le loro origini e modalità d'azione si trovavano al centro di un dibattito tra diverse teorie.

La formazione degli anticorpi pone numerosi problemi a causa della diversità quasi incredibile di antigeni, le sostanze estranee che provocano una risposta immunitaria. Un anticorpo può neutralizzare solo un tipo di antigene, ma gli antigeni sono estremamente variabili per forma, dimensioni e composizione chimica. I batteri e le loro tossine, i virus, i granuli di polline, le cellule del sangue incompatibili e le molecole di sintesi agiscono tutti come antigeni. Per catturare ogni tipo di invasore, i linfociti *B*, particolari globuli bianchi del sangue, devono produrre grandi quantità di anticorpi «su misura».

Il punto principale della contesa tra gli scienziati che per primi svilupparono queste teorie, schierandosi ben presto in due campi, fu il modo in cui gli antigeni regolano la risposta immunitaria. Una scuola di pensiero riteneva che gli antigeni servissero da stampo per dirigere la progettazione di anticorpi che si adattassero a essi. L'altra scuola riteneva, invece, che i linfociti conservassero una scorta di anticorpi già pronti, in mezzo ai quali un antigene sceglieva quello che gli si adattava meglio. Negli ultimi trent'anni, molti ricercatori in vari paesi hanno unito i loro sforzi e sono riusciti a chiarire la base biologica dell'immunità, risolvendo la controversia. Ha vinto il secondo punto di vista, sviluppatosi poi nella teoria della selezione clonale per la formazione degli anticorpi.

Questo articolo espone le vicende attraverso le quali questa teoria è stata sviluppata e convalidata. Fondamentalmente essa fu proposta quasi un secolo fa, ma subì in seguito alterne vicissitudini, cadendo in disgrazia per poi essere di nuovo riabilitata, in base a prove concrete o a semplici congetture. Entrambi gli autori hanno avuto il privilegio di prendere parte alla ricerca che, negli anni sessanta, ha dato alla teoria il suo giusto riconoscimento.

Le ricerche sul meccanismo di formazione degli anticorpi ebbero inizio alla fine dell'Ottocento, periodo in cui la teoria di Louis Pasteur, secondo la quale gli agenti causali delle malattie sono germi, era ormai universalmente accettata. Diversi gruppi cominciarono a studiare la reazione tra tossine batteriche e «antitossine», che compaiono nel siero, cioè nella componente liquida del sangue, dopo l'infezione. Il batteriologo tedesco Emil A. von Behring coniò per queste sostanze il termine *Antikörper*, ossia anticorpi.

Le reazioni tra anticorpi e antigeni potevano essere osservate in provetta, dato che i reagenti formavano aggregati visibili a occhio nudo. Fu presto evidente che i prodotti batterici non erano l'unico agente in grado di scatenare la produzione di anticorpi. Altre sostanze naturali, come le proteine presenti nel latte e le cellule «estranee», potevano attivare la risposta immunitaria da parte dell'«ospite».

Nel 1890 von Behring, che aveva prodotto un'antitossina contro la difterite, conobbe Paul Ehrlich, un altro medico tedesco, e questi, in seguito, pubblicò un articolo fondamentale, in cui veniva descritta una tecnica per misurare gli anticorpi contro la difterite in preparati come quelli utilizzati da von Behring. La tecnica permetteva la standardizzazione dei preparati e, di conseguenza, rendeva sicuro per la pratica clinica il siero antidifterico.

Nel mettere a punto la tecnica per misurare gli anticorpi, Ehrlich stabilì le basi per affrontare quantitativamente l'immunologia. Le nuove osservazioni quantitative rivelarono che la risposta immunitaria consisteva in una proliferazione esplosiva di anticorpi in seguito al contatto con un agente infettivo. Per spiegare come le sostanze estranee potessero indurre questa reazione, Ehrlich sviluppò la teoria delle catene laterali per la formazione degli anticorpi.

Enunciata al volgere del secolo, la teoria delle catene laterali postulava che la superficie di un globulo bianco presentasse recettori con catene laterali alle quali si legavano chimicamente sostanze estranee. Questo legame induceva la cellula a produrre in grande eccesso copie del recettore legato. I recettori superflui, che erano gli anticorpi, si staccavano e venivano riversati nel sangue. Ehrlich ebbe il merito di formulare l'ipotesi che i globuli bianchi fossero in grado di produrre spontaneamente le catene laterali capaci di legare tutte le sostanze estranee.

Alcuni critici della teoria di Ehrlich si opposero all'idea di una unione chimica tra l'antigene e il suo recettore. Altri ricercatori accettarono il concetto, ma misero in discussione il meccanismo di legame dell'antigene. La cellula «fagocitava» il complesso antigene-recettore, oppure la reazione era reversibile, con i recettori che catturavano gli antigeni e poi li lasciavano andare? In mezzo a queste dispute, lo studio delle reazioni antigene-anticorpo e della specificità degli anticorpi assunse sempre maggiore importanza. Si sapeva che gli anticorpi cominciano a circolare nel sangue subito dopo il primo contatto con un antigene e che un secondo contatto con lo stesso antigene dà luogo a una risposta più rapida e più potente. Nessuno sapeva, pe-

Il medico tedesco Paul Ehrlich può essere ritenuto il padre della teoria della selezione clonale. Nel 1900 avanzò l'ipotesi che gli anticorpi fossero presenti sulle superfici cellulari come recettori specifici. Contribuì, inoltre, a sviluppare un antisiero per la cura della difterite e a sintetizzare un farmaco per la cura della sifilide. Nel 1908 ricevette il premio Nobel.

rò, fino a che punto fossero particolari questi anticorpi.

Come è stata sviluppata la teoria dello stampo

Molti ricercatori hanno affrontato il problema della specificità degli anticorpi esaminando gli effetti di antigeni modificati. Nel 1906, in Germania, Ernest P. Pick e Friedrich P. Obermayer dimostrarono che, attaccando a una proteina gruppi chimici come lo iodio o il nitrato, se ne modificano profondamente le proprietà antigeniche. All'incirca nello stesso periodo, l'immunologo austriaco Karl Landsteiner cominciò a seguire una analoga strategia sperimentale. Il suo lavoro, svolto per la maggior parte al Rockefeller Institute for Medical Research, continuò per ben un trentennio e si concluse dimostrando l'assoluta specificità degli anticorpi.

Landsteiner mise in contatto le proteine antigeniche di cui disponeva con una grande varietà di gruppi chimici, alcuni dei quali provenivano da microbi patogeni e altri erano sintetizzati in provetta. Ogni molecola alterata induceva la formazione di un diverso anticorpo. Pertanto, Landsteiner confermò l'osservazione di Pick e Obermayer che le subunità chimiche di una struttura antigenica più grossa potevano determinarne la specificità immunologica. Questi determinanti, vale a dire semplici strutture molecolari estranee, non si comportavano come antigeni veri e propri, cioè non potevano avviare la produzione di anticorpi, o formare aggregati con gli anticorpi, a meno che non fossero legati a una molecola vettore.

Oggi sappiamo che molti determinanti non possono indurre una risposta immunologica da soli perché sono troppo piccoli; per poter essere riconosciuti devono pertanto attaccarsi a una molecola di maggiori dimensioni. Ottant'anni fa, però, la constatazione che i determinanti non fossero antigeni lasciò piuttosto sorpresi. Inoltre, la scoperta dei determinanti antigenici indicava che la gamma di sostanze estranee che potevano scatenare la risposta del sistema immunitario era molto maggiore di quanto si sospettasse. Oggi è ormai assodato che un singolo agente invasore, per esempio un batterio, potrebbe portare migliaia di determinanti e avviare la produzione di migliaia di differenti anticorpi.

Dagli studi di Landsteiner risultò chiaro che un animale era in grado di sintetizzare una gamma praticamente illimitata di anticorpi, anche diretti contro nuovi composti artificiali. Queste scoperte portarono a concludere che un ospite non poteva intrinsecamente possedere l'informazione di cui si serviva per rispondere a uno spettro così ampio di antigeni. La teoria delle catene laterali di Ehrlich fu, pertanto, scartata in favore dell'idea che gli antigeni potessero in certo qual modo dirigere la specificità degli anticorpi, quando questi venivano sintetizzati all'interno delle cellule sanguigne.

In che modo un antigene avrebbe potuto influire sulla conformazione di un anticorpo? Agli inizi degli anni trenta furono avanzate numerose proposte, tutte necessariamente frutto di ipotesi. Friedrich Breinl e Felix Haurowitz dell'Università di Praga e Stuart Mudd e Jerome Alexander dell'Università della Pennsylvania ritenevano che gli anticorpi venissero sintetizzati a diretto contatto con i loro antigeni, assumendo forma e affinità chimica complementari a quelle del rispettivo antigene. Essi suggerivano che gli anticorpi modellati da antigeni diversi differissero nella composizione delle proteine: avessero, cioè, diverse sequenze delle unità costitutive, gli amminoacidi. Altri ricercatori, in particolare Linus Pauling, sostenevano che un ampio ventaglio di specificità di legame potesse derivare dal solo ripiegamento della stessa molecola proteica dell'anticorpo in differenti modi. Anche in questo caso, il ripiegamento sarebbe stato guidato dallo stampo antigenico. Tutte queste ipotesi, collettivamente, sono confluite nella teoria dello stampo per la formazione degli anticorpi, una teoria che ha dominato il campo per un quarto di secolo.

Queste discussioni sui meccanismi di formazione degli anticorpi si svolsero prima dell'avvento della biologia molecolare negli anni cinquanta; esse, quindi, si basavano più sulla teoria che su dati sperimentali. La prima opposizione seria alla teoria dello stampo venne perciò dai biologi e non dai biochimici. Tre scienziati, altrettanto degni di considerazione di qualunque altro li avesse preceduti, raccolsero le prove decisive. Due di essi avrebbero in seguito ricevuto il premio Nobel, come già era accaduto a von Behring, Ehrlich e Landsteiner.

I limiti della teoria dello stampo

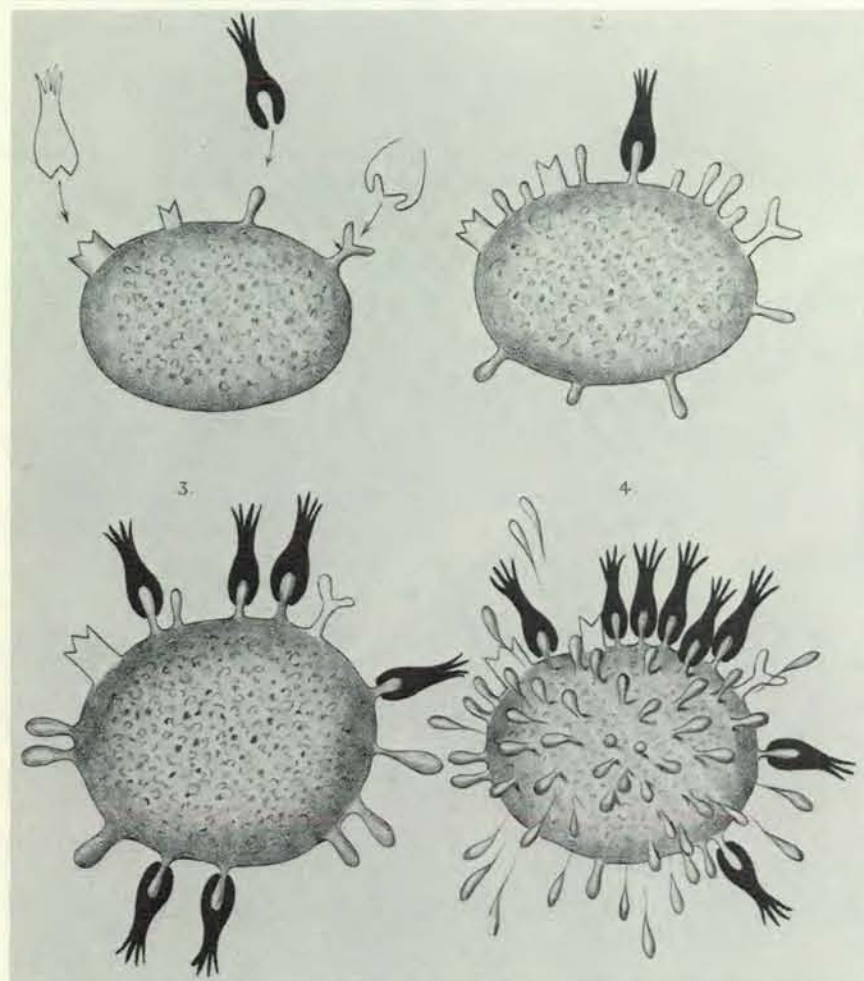
Il primo attacco globale a questa teoria fu sferrato nel 1955 dall'immunologo danese Niels Kaj Jerne. In un articolo pubblicato in quell'anno, Jerne fece presente che, nel 1949, due ricercatori australiani, F. Macfarlane Burnet e Frank J. Fenner, avevano fatto diverse osservazioni che non si conciliavano con la teoria dello stampo. In primo luogo, la teoria non riusciva a spiegare l'aumento apparentemente esponenziale nella produzione di anticorpi durante i primi stadi di una risposta immunitaria. Se per produrre ciascun anticorpo era necessario uno stampo antigenico, era difficile immaginare come gli anticorpi potessero così rapidamente superare di numero i rispettivi stampi. Inoltre, la teoria non spiegava il forte incremento di anticorpi che si notava quando un animale si trovava in presenza di un determinato antigene per la seconda volta. Perché ripresentando uno stampo si dovrebbero avere più copie dell'anticorpo, rispetto al contatto iniziale?

Anche la produzione di anticorpi, che prosegue per molto tempo dopo la scomparsa dell'antigene, creò problemi alla teoria dello stampo, poiché si pensava che le cellule produttrici di anticorpi avessero vita breve. Inoltre, questa teoria non prendeva in considerazione il fatto che, a mano a mano che la reazione immunitaria proseguiva, gli anticorpi sembravano migliorare il loro legame con gli antigeni bersaglio.

Forse la sfida più avvincente che Burnet e Fenner raccolsero fu il fenomeno della tolleranza immunologica. Questa, che è in pratica l'incapacità di sferrare un attacco immunologico contro un dato antigene, era già stata riconosciuta non molto tempo prima che Burnet e Fenner pubblicassero le loro critiche. La tolleranza impedisce a un animale di produrre anticorpi contro se stesso e può essere acquisita per antigeni estranei, se questi vengono somministrati prima, o al momento, della nascita. Contrariamente all'immunità, la tolleranza non può essere mantenuta dall'organismo, a meno che l'antigene non persista in esso. Sull'argomento della tollerenza immunologica, la teoria dello stampo non prendeva posizione.

Jerne lanciò queste critiche nell'articolo *The Natural-Selection Theory of Antibody Formation* per poter formulare un concetto che sarebbe sembrato familiare a Ehrlich e cioè che qualunque animale possiede piccole quantità di anticorpi contro tutti gli antigeni. Si realizza una risposta immunitaria quando un antigene si lega a un anticorpo e il complesso antigene-anticorpo interagisce con i globuli bianchi, stimolando la produzione e la liberazione in grandi quantità dello stesso anticorpo specifico. La proposta di Jerne era suffragata dall'osservazione che il siero di sangue normale contiene sempre globuline, anticorpi non specifici, che a quell'epoca sembravano differenziarsi dagli altri anticorpi solo per la loro mancanza di specificità. Jerne citava anche una rassegna scritta da Robert Doerr dell'Università di Basilea, in cui venivano riassunte le prove in favore degli anticorpi «naturali», anticorpi che erano prodotti senza uno stimolo antigenico.

Alla School of Medicine dell'Università del Colorado a Denver, David W. Talmage lesse l'articolo di Jerne e rilevò le similitudini tra il modo di pensare di questo autore e quello espresso da Ehrlich con la sua teoria delle catene laterali. Egli andò poi un poco più in là. In un articolo del 1957, suggerì che la risposta immunitaria doveva essere caratterizzata sia da cellule che si duplicavano, sia da molecole di anticorpo liberamente circolanti nel sangue. Egli sosteneva che le cellule vengono selezionate per la moltiplicazione quando l'anticorpo che esse sintetizzano si adatta perfettamente all'antigene invasore. Talmage fece anche rilevare che una cellula cancerosa che produce anticorpi dà origine a un flusso notevolmente omogeneo di molecole, indicando così che le singole cellule potrebbero «specializzarsi» nella sintesi di un particolare anticorpo.

In retrospettiva, appare evidente che Jerne e Talmage, nei loro due articoli, avevano posto le basi della teoria della selezione clonale per la formazione degli anticorpi. Rimaneva, però, a Burnet il compito di realizzare una sintesi organica dei nuovi concetti sviluppati attraver-

L'illustrazione, relativa alla teoria delle catene laterali, compariva nel lavoro pubblicato da Ehrlich nel 1900 in cui era enunciata l'idea che una sostanza estranea (*in nero*) si legasse a un recettore della cellula (*1, 2*), stimolando quest'ultima a produrre e liberare gli anticorpi (*3, 4*).

Secondo la teoria della selezione clonale, che negli anni sessanta acquistò considerazione tra i ricercatori, le cellule produttrici di anticorpi hanno recettori specifici e ogni cellula ne costruisce un solo tipo. Se un antigene si adatta bene a un recettore, si lega a esso (*1*) e induce la cellula a dividersi e a produrre altri recettori (*2*). Come nel modello di Ehrlich, i recettori si staccano dalla cellula ed entrano in circolo (*3*).

so le sue riflessioni con Fenner. Burnet aveva fatto notevoli sforzi per definire il meccanismo per mezzo del quale il contatto di un antigene con una cellula avrebbe innescato un sistema di autoreplicazione. Come prima cosa, egli immaginò che gli antigeni avessero la funzione di «istruire» gli enzimi adattativi, e che questi ultimi, una volta imparata la configurazione dell'antigene, potessero procedere nella sintesi dell'anticorpo anche in assenza del loro istruttore. In seguito, egli avanzò l'ipotesi che gli antigeni interagissero direttamente con il materiale genetico di una cellula. Queste ipotesi di lavoro non sono più valide, ma permisero comunque a Burnet di sottolineare l'importanza del ruolo svolto dalla cellula e dalla sua replicazione nella produzione di anticorpi.

La selezione clonale

L'articolo di Jerne fornì a Burnet l'anello mancante, vale a dire l'idea che l'organismo fosse dotato di anticorpi preesistenti per riconoscere tutti gli antigeni. Sulle orme di Ehrlich e Talmage, Burnet propose che il legame di un antigene con un recettore a cui è fissato un anticorpo innescasse la moltiplicazione cellulare e la produzione da parte della cellula di un maggior quantitativo dello stesso recettore. Quindi, imboccando una diramazione del filone pricipale del suo ragionamento, egli asserì che ogni cellula e i suoi cloni, cioè la discendenza di quella cellula, potevano produrre un solo tipo di recettore. Per descrivere questa teoria, Burnet usò appunto il termine «selezione clonale».

La teoria della selezione clonale affascinava Burnet perché dava una risposta a diverse critiche alla teoria dello stampo che egli aveva espresso in precedenza. Secondo la teoria, la crescita esponenziale nella produzione di anticorpi per contatto con l'antigene è dovuta all'aumento esponenziale nel numero di cellule che producono quegli anticorpi. La reazione secondaria a un antigene è più potente e più rapida della prima perché vi è un maggior numero di cellule che reagisce dopo la stimolazione iniziale dell'antigene. Una volta completata l'«équipe» delle cellule che producono un particolare anticorpo, non è più necessaria un'esposizione prolungata all'antigene per mantenere la produzione di anticorpi. Con il passare del tempo gli anticorpi migliorano la loro capacità di formare legami perché l'antigene «seleziona» per la duplicazione quelle cellule con mutazioni genetiche che promuovono un buon abbinamento tra anticorpo e antigene. Infine, la teoria della selezione clonale spiega la tolleranza immunologica come delezione di un intero clone di cellule, che può verificarsi prima o subito dopo la nascita, o anche in seguito, se un antigene è in grado di annientare le capacità metaboliche di quelle cellule.

Burnet concepiva la risposta immunitaria come una specie di microcosmo darwiniano. Le cellule produttrici di anticorpi, come un qualunque organismo in un ecosistema, sono soggette a mutazione e a selezione; le più idonee sopravvivono e l'idoneità, in questo caso, è letteralmente l'adattamento tra un anticorpo prodotto da una cellula dell'organismo e l'antigene. Nel 1957, dopo aver preso visione della rassegna di Talmage, Burnet propose il suo manoscritto *The Clonal Selection Theory of Antibody Formation* a una rivista quasi sconosciuta. Forse egli era consapevole delle imperfezioni contenute nei concetti che aveva precedentemente proposti e solo nel 1958 cominciò a scrivere un libro in cui elaborava la propria teoria, mentre ancora più tardi mise per iscritto, su riviste più importanti, le idee che andava mano a mano sviluppando.

Nel decennio successivo, Burnet elaborò e portò avanti con grande pervicacia la teoria della selezione clonale. Il premio Nobel che lui e P. B. Medawar condivisero nel 1960 riconosceva i suoi successi nella comprensione della tolleranza immunologica acquisita, ma Burnet riteneva che la sua realizzazione più significativa consistesse, invece, nel lavoro compiuto per promuovere la teoria della selezione clonale. Nel 1984 anche Jerne ricevette il premio Nobel per i suoi contributi teorici all'immunologia, il più importante dei quali fu senza dubbio il ruolo che egli svolse nello sviluppo della teoria selettiva.

Una cellula, un anticorpo?

Nel febbraio del 1957, uno di noi (Nossal) aveva iniziato una ricerca come borsista presso il laboratorio di Burnet, prefiggendosi di esaminare i problemi della tolleranza immunologica nei feti di topo. Quando, quello stesso anno, Burnet annunciò la teoria della selezione clonale, parve naturale suggerire, nello stile popperiano, un esperimento che la falsificasse.

Come ogni allievo di Burnet, Nossal fu stimolato a leggere la letteratura corrente sui virus e fu colpito dalla notizia, più volte riportata, che le cellule utilizzate per coltivare i virus animali potevano essere isolate e mantenute in vita in colture di singole cellule. La programmazione degli esperimenti finì per condurlo a verificare la teoria della selezione clonale. L'idea di Burnet era che una cellula fosse in grado di produrre un solo tipo di anticorpo. Se fosse stato possibile isolare una singola cellula produttrice di anticorpi, sarebbe anche stato possibile determinare se quella cellula produceva più di un solo tipo di anticorpo.

Per una fortunata coincidenza, Joshua Lederberg, che allora lavorava all'Università del Wisconsin a Madison, stava anch'egli trascorrendo alcuni mesi nel laboratorio di Burnet con una borsa di studio Fulbright. Lederberg, uno dei padri della genetica batterica, aveva una esperienza considerevole nell'inserire singole cellule batteriche in minuscole goccioline di brodo colturale per esaminarle poi al microscopio ottico. Egli mise spontaneamente a disposizione le sue capacità per realizzare la sperimentazione necessaria per la verifica della teoria della selezione clonale, il che consentì di utilizzare il moto batterico come indicatore della produzione di anticorpi.

Si sapeva che i batteri potevano venire completamente immobilizzati lungo le loro abituali traiettorie, se messi in presenza degli anticorpi contro i loro flagelli, le delicate strutture filiformi che svolgono una funzione propulsiva. Nossal e Lederberg immunizzarono alcuni ratti con due differenti antigeni flagellari e, alcuni giorni dopo, ne asportarono i linfonodi, gli organi che, insieme alla milza, costituiscono una sede importante di produzione di anticorpi.

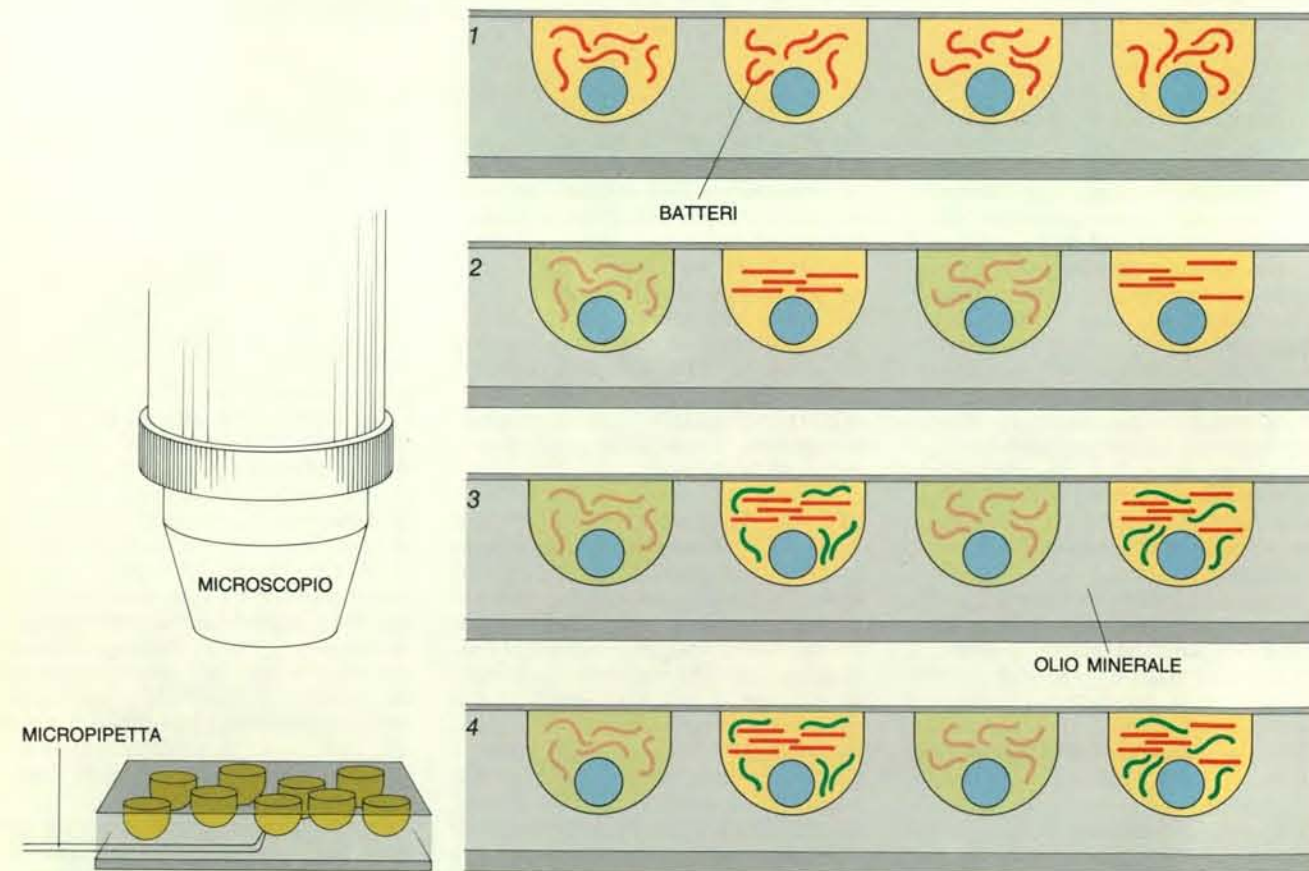
Servendosi di aghi molto sottili, essi cercarono di frammentare il tessuto in modo da ottenere una sospensione di cellule singole; introdussero quindi queste cellule, una per una, in goccioline di volume non maggiore di un milionesimo di millilitro. Per impedire alle goccioline di evaporare, le circondarono con uno straterello di olio minerale e, dopo una breve incubazione, a ciascuna aggiunsero dai cinque ai dieci batteri in rapido movimento e con flagelli identici. Compirono, quindi, un'osservazione al microscopio. Se i batteri apparivano immobilizzati, indicando così la presenza di un tipo di anticorpo, inserivano i batteri con l'altro tipo di flagello.

I ricercatori arrivarono così alla conclusione che, mentre erano numerose le cellule in grado di produrre anticorpi contro un unico tipo di flagello, nessuna cellula produceva anticorpi contro entrambi i tipi. Così, la regola «una cellula, un anticorpo» fu stabilita proprio grazie a un esperimento che era stato progettato per invalidarla.

Reazioni crociate disarmanti

Nel 1959 Nossal raggiunse Lederberg alla School of Medicine della Stanford University per installarvi un nuovo laboratorio di immunologia. Qui egli iniziò una collaborazione biennale con Olavi Mäkelä dell'Università di Helsinki, che confermò e ampliò le conclusioni del gruppo di Burnet. Questa collaborazione stimolò la fantasia di Mäkelä, il quale, al suo ritorno a Helsinki, cominciò a studiare attentamente le reazioni immunologiche crociate, un fenomeno attraverso il quale gli anticorpi prodotti contro un antigene reagiscono anche contro altri antigeni simili. Per le sue indagini, egli decise di adoperare colture costituite da un unico tipo di cellula.

Mäkelä provocò la formazione di anticorpi servendosi di virus batterici, o fagi, e quantificò la risposta che la cellula dà con la produzione di anticorpi, misurando quanti fagi una cellula poteva rendere innocui, cioè incapaci di infettare i batteri loro ospiti. Prima di esaminare gli anticorpi di singole cellule, egli misurò la riserva di anticorpi presenti nel siero sanguigno di un animale immunizzato. In questo caso trovò esattamente quello che i libri di testo gli avevano insegnato che avrebbe dovuto trovare: gli animali immunizzati con un fago immunogeno *A* producono anticorpi che neutralizzano molto bene questo fago, ma che hanno anche una reazione crociata con il fago *B*, con un'efficienza, però, solo del 20 per cento.

Quando Mäkelä osservò gli anticorpi nelle goccioline contenenti una singola cellula, ottenne risultati nettamente diversi. Gli anticorpi di ogni cellula avevano una differente specificità per il fago *A*. La maggior parte di essi, ma non tutti, neutralizzava il fago *A* meglio del fago *B*. Sembrava che ogni cellula producesse anticorpi dalle proprietà leggermente

Le colture monocellulari hanno permesso agli immunologi di verificare quanti anticorpi diversi possa produrre una cellula. Alcuni ratti sono stati immunizzati contro due batteri mobili (*in rosso e in verde*); poi, dai loro linfonodi, sono state isolate cellule che, poste in gocce di liquido colturale osservabili al microscopio (*a sinistra*), sono state messe in contatto con i batteri (*1*). Nelle cellule capaci di immobilizzare un tipo di batterio (*2*), veniva iniettato anche il secondo tipo (*3*). Molte cellule potevano arrestare un tipo di batterio o l'altro; nessuna entrambi (*4*).

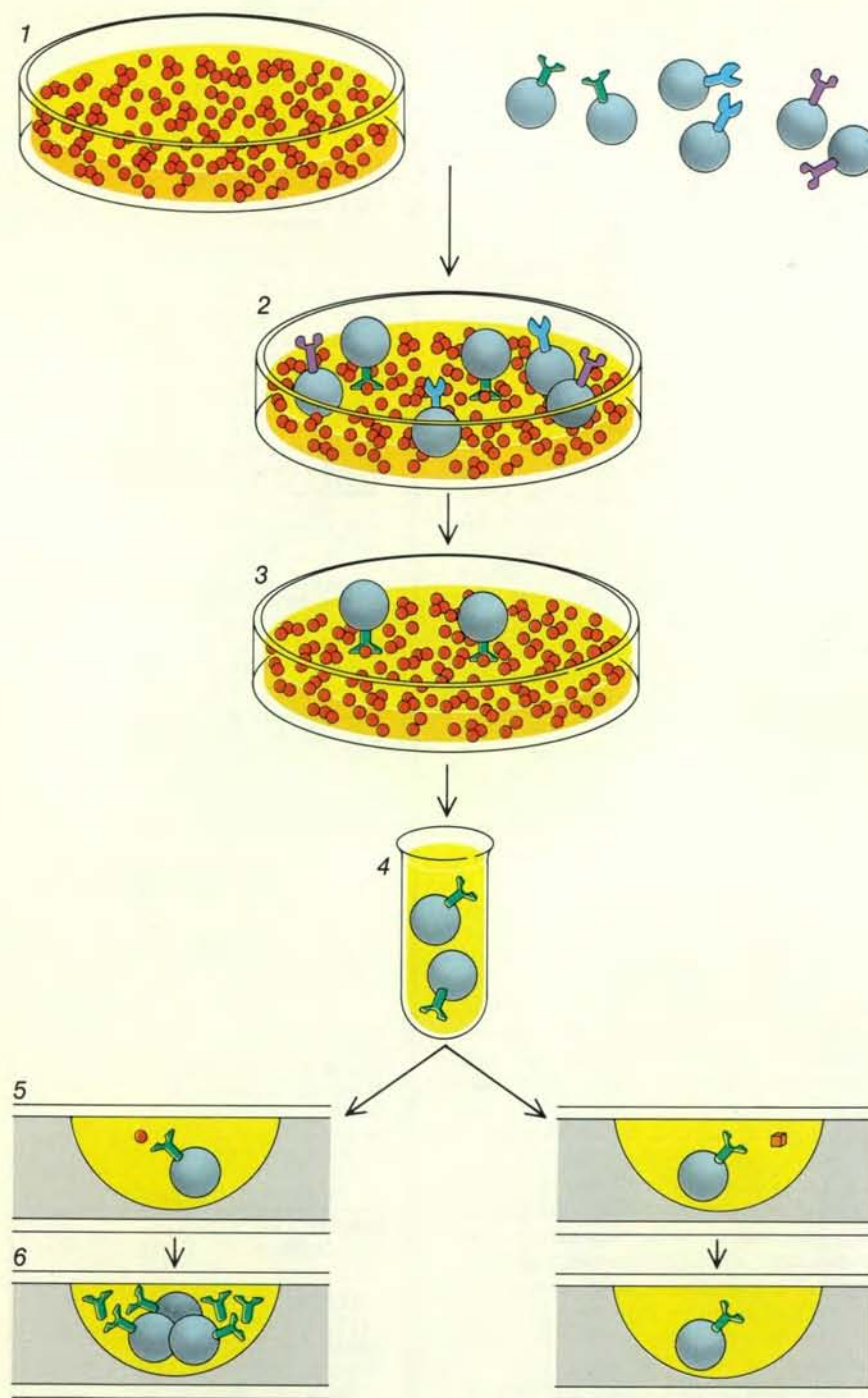
diverse. Le reazioni che Mäkelä aveva osservato nel siero rispecchiavano semplicemente l'effetto cumulativo di molte cellule peculiari, ciascuna delle quali «svolgeva una propria attività». Questa scoperta era esattamente quanto veniva previsto dalla teoria della selezione clonale.

La teoria dello stampo decade

Per quasi cinque anni dopo gli esperimenti sopra descritti, lo studio della formazione di anticorpi da parte di singole cellule fu praticamente una riserva privata dello stesso Nossal, di Mäkelä e di chi aveva svolto ricerche con loro. Poi, nel 1963, Jerne e Albert A. Nordin dell'Università di Pittsburgh svilupparono un metodo molto più semplice per individuare la presenza di anticorpi secreti da singole cellule. Le cellule della milza e dei linfonodi di un topo immunizzato con globuli rossi di pecora venivano distese su una piastra di gelatina rivestita di cellule ematiche. Quando venivano aggiunte determinate proteine, attorno a ogni cellula che produce anticorpi compariva una zona chiara. Questa tecnica si è dimostrata flessibile e accurata per il conteggio delle cellule attive e la sua disponibilità ha indotto molti altri laboratori a studiare le questioni sollevate dalla teoria di Burnet.

Mentre altri ricercatori si dedicavano allo studio di quanto era stato previsto dalla teoria della selezione clonale, nel 1962 demmo il via alla nostra quinquennale collaborazione con un esperimento che speravamo avrebbe chiarito una volta per tutte la validità della teoria dello stampo. Desideravamo seguire il percorso delle molecole di antigeni nell'organismo. A quell'epoca era noto che gli anticorpi e tutte le altre proteine venivano sintetizzati da strutture specializzate, i poliribosomi. Se la teoria dello stampo fosse stata giusta, ogni poliribosoma avrebbe dovuto essere associato a una molecola di antigene o, perlomeno, a un suo frammento significativo. Dal momento che ogni cellula è dotata di migliaia di poliribosomi attivi, è probabile che ogni cellula in grado di produrre anticorpi debba contenere molte copie dell'antigene o di suoi frammenti.

Anche in questo caso gli antigeni dei flagelli batterici ci servirono come strumenti di ricerca. Mettemmo in contatto un isotopo radioattivo dello iodio con un antigene flagellare, in modo da poter localizzare l'antigene e i suoi frammenti «marcati» ricercando la radioattività in cellule o in campioni di tessuti prelevati dall'animale trattato. Non una sola cellula produttrice di anticorpi conteneva una quantità misurabile di antigene marcato, anche se i nostri metodi erano abbastanza sensibili da consentirci di trovare 10 molecole di antigene per cellula. Scoprimmo, invece, che la maggior parte dell'antigene era finita nei macrofagi, cellule che potremmo definire «spazzine». Questa prova era incompatibile con le previsioni della teoria dello stampo.

Nonostante il largo uso da noi fatto di antigeni marcati per rintracciare le cellule immunologicamente attive, non avevamo tentato di trovare anticorpi in animali non immunizzati. Eppure una delle asserzioni più controverse della teoria della selezione clonale era appunto che un piccolo numero di recettori naturali (anticorpi) per tutti gli antigeni fosse presente nell'organismo prima dell'immunizzazione. Si sarebbe potuto ribattere che tale numero era così piccolo che i recettori potevano facilmente sfuggire all'individuazione, con il risultato che alcune globuline del siero sangui-

L'isolamento delle cellule che legano gli antigeni è servito a dimostrare che esse danno origine a cellule che producono anticorpi. Milioni di cellule di milza non immunizzate vengono immesse in una piastra dove è stata versata gelatina contenente molecole antigeniche identiche (*1*). Le cellule munite di recettori specifici per quell'antigene aderiscono a esso (*2*); le altre sono asportate (*3*). Dopo che la gelatina è stata fusa e le cellule sono state disperse con l'aiuto di enzimi (*4*), ogni cellula viene posta nella propria coltura (*5*). Alcune cellule sono esposte all'antigene originale (*a sinistra*) mentre altre sono messe in contatto con antigeni che non hanno relazione con esso (*a destra*). Le cellule trattate con l'antigene originale si duplicano e producono anticorpi (*6*).

Negli anni cinquanta tre scienziati hanno ridato vitalità alla teoria della selezione clonale. N. K. Jerne (*in alto a sinistra*) ha proposto che l'antigene agisca come agente selettivo; D. W. Talmage (*a destra*) ha suggerito che il legame con l'antigene induca le cellule a moltiplicarsi; F. M. Burnet (*in basso a sinistra*) ha sostenuto che ogni cellula produce un solo tipo di anticorpo.

gno sembravano essere non specifiche.

Due ricercatori dell'Università ebraica di Gerusalemme, David Naor e Dov Sulitzeanu, ritenevano che il segnale proveniente dagli antigeni radioattivi legati ai recettori potesse essere abbastanza forte da venire individuato. Essi esposero cellule di milza di topi non immunizzati a un antigene marcato e cercarono la prova che questo si fosse legato. In base alla teoria della selezione clonale, essi si aspettavano che le cellule dotate di recettori per l'antigene dovessero essere in numero ristretto e quindi si attendevano di trovare tracce di radioattività su pochissime cellule. Questa previsione si verificò puntualmente. Inoltre, come era previsto dalla teoria della selezione clonale, gli animali immunizzati avevano un numero di cellule in grado di legare l'antigene superiore a quello degli animali non immunizzati, mentre gli animali tolleranti ne avevano di meno. Naor e Sulitzeanu completarono il loro lavoro sperimentale nel 1967 e la loro ricerca sollevò immediatamente un interrogativo fondamentale: come si poteva essere sicuri che le cellule che legano gli antigeni fossero veramente i predecessori delle cellule che producono gli anticorpi?

Il suicidio degli antigeni

Uno di noi (Ada), in collaborazione con Pauline Byrt del Walter and Eliza Hall Institute of Medical Research di Melbourne, propose un metodo di laboratorio per rispondere a quella domanda. Se le cellule produttrici di anticorpi sono derivate, di fatto, esclusivamente dalle cellule che legano gli antigeni, allora il danno prodotto alle loro cellule progenitrici da un antigene radioattivo dovrebbe impedire la replicazione delle cellule che producono gli anticorpi specifici per quegli antigeni. Altre cellule non sarebbero colpite.

Nel 1969, la Byrt e Ada cominciarono a occuparsi degli esperimenti detti, con termine pittoresco, «suicidio degli antigeni caldi», dove l'aggettivo caldo indica gli antigeni marcati. Due antigeni di *Salmonella*, distinti ma simili, ciascuno legato a iodio radioattivo, vennero iniettati in due gruppi separati di topi. I linfociti degli animali furono quindi trasferiti a due altri gruppi di topi, i cui sistemi immunitari erano stati distrutti tramite raggi X. Quindi i due gruppi di topi vennero inoculati con entrambi gli antigeni, privi questa volta della loro marcatura radioattiva. Nei topi irradiati, i sistemi immunitari sostitutivi secernevano pochi anticorpi, o niente affatto, in risposta all'antigene a cui i loro linfociti erano stati esposti, mentre davano la solita energica risposta all'altro antigene. In altre parole, quando le cellule che legano gli antigeni erano danneggiate, anche la risposta immunitaria all'antigene risultava deficitaria.

Gli esperimenti di suicidio degli antigeni marcati diedero alla teoria della selezione clonale un impulso significativo. Nel frattempo Hans L. R. Wigzell e Birger Andersson del Karolinska Institutet di Stoccolma erano giunti ad analoghe conclusioni ricorrendo a una strategia alternativa. Tuttavia, nessuno degli esperimenti condotti fino ad allora poteva dimostrare in modo inequivocabile che la superficie cellulare ha un unico tipo di recettore. Era chiaro che i prodotti di una singola cellula erano meno eterogenei degli anticorpi presenti nel siero, ma era ancora possibile che sulla superficie di una singola cellula fossero rappresentate 10 o 100 specificità, al posto di diverse migliaia.

Nel 1973, è stato effettuato un esperimento ingegnoso, progettato da Martin C. Raff, Marc Feldman e Stefanello de Petris dell'University College di Londra, nel tentativo di eliminare quest'incertezza. In precedenza, Raff aveva scoperto una sostanza chimica che poteva formare legami crociati con i recettori di superficie presenti sulle cellule. Tale sostanza era dotata di azione irritante sulla cellula, con il risultato che tutti i recettori legati a essa venivano circoscritti in un settore della membrana cellulare ben delimitato e a forma di calotta. Una sostanza in grado di formare legami crociati e di agire con la specificità di un antigene riunirebbe in un settore della cellula solo quei recettori ai quali è legata. I ricercatori avevano a disposizione un antigene che forse rispondeva alle loro esigenze. Sarebbe riuscito a legare e riunire in un'unica struttura tutti i recettori di una data cellula, oppure ne avrebbe lasciati da parte alcuni? Raff e collaboratori lo hanno messo alla prova con i globuli bianchi del sangue e hanno trovato che, sulle cellule che reagivano, più del 95 per cento dei recettori erano stati legati e riuniti in una struttura a calotta.

La dimostrazione definitiva

Ora era necessaria soltanto una dimostrazione formale per suggellare il successo della teoria della selezione clonale. Il rigore scientifico esigeva che venisse effettuato un esperimento in grado di seguire una popolazione omogenea di cellule che legano gli antigeni dalla stimolazione da parte di un antigene alla produzione di anticorpi. Tuttavia, fino agli inizi degli anni settanta, i tentativi per preparare una popolazione di cellule sane che reagissero tutte a un unico antigene furono frustranti e infruttuosi. Alla fine, quattro o cinque gruppi di ricercatori, tra cui quello diretto da Nossal, riuscirono a mettere a punto alcune tecniche promettenti.

Il gruppo di Nossal propose un metodo semplice, che sfruttava le proprietà di fusione della gelatina. I ricercatori aggiungevano alla gelatina in forma liquida molecole di un dato antigene e lasciavano solidificare la miscela antigene-gelatina in piastre poco profonde. Quindi immettevano in ogni piastra 100 milioni di cellule normali di milza provenienti da un animale non immunizzato. Le cellule specifiche per l'antigene incluso nella gelatina avrebbero aderito a esso, consentendo così di distinguerle dalle altre cellule che avrebbero potuto venire scartate. In questo modo si riuscivano a isolare nella gelatina le cellule che reagivano con un solo antigene.

Facendo fondere la gelatina, i ricercatori liberavano le cellule specifiche per l'antigene. Questo, a sua volta, era liberato per mezzo di un enzima che digerisce la gelatina e trascina con sé l'antigene. Una volta in soluzione, ogni cellula veniva posta in una propria coltura per mezzo di tecniche messe a punto in gran parte da Beverley L. Pike del Walter and Eliza Hall Institute. Ogni cellula veniva poi stimolata con l'antigene originale e con un secondo antigene estraneo al primo. Come si aspettavano, i collaboratori di Nossal trovarono che cellule singole davano origine a cloni produttori di anticorpi in risposta al primo antigene e non all'altro. Infatti il mezzo di coltura che circondava i cloni era ricco di anticorpi contro il primo antigene, mentre non ne conteneva contro il secondo.

Lo scopo della selezione

Questi esperimenti di clonazione erano difficilmente realizzabili quando se ne parlò la prima volta nel 1976, ma da allora sono diventati molto più fattibili e i risultati originali sono stati riprodotti parecchie volte. Oggi il concetto di selezione clonale viene accettato come realtà ed è stata rigorosamente confermata la derivazione diretta di cellule che producono gli anticorpi da cellule che legano gli antigeni. Queste ultime sono i linfociti *B*; le loro discendenti, le cellule clonali che producono anticorpi, sono note come plasmacellule.

Questa nostra rassegna può dare l'impressione che ci sia stata una progressione semplice, diretta, dalla formulazione della teoria alla sua convalida. In realtà il percorso è stato molto più complicato. Parecchi esperimenti che non abbiamo menzionato hanno avuto un impatto incidentale, ma significativo. Anche le tecniche per il trasferimento di sistemi immunitari sostitutivi ai topi irradiati con raggi X, la comprensione della struttura della molecola dell'anticorpo e la scoperta della base genetica della diversità degli anticorpi hanno contribuito al successo della teoria della selezione clonale.

L'ultimo apporto risponde a una domanda particolarmente spinosa che ha assillato chi si dedicava alla teoria della selezione clonale: come possono i linfociti prevedere la grande varietà di antigeni e possedere recettori in grado di intercettarli tutti? Lo stesso Burnet riconobbe che la sua teoria richiedeva un certo tipo di aleatorietà per il quale non c'erano precedenti. Si sa oggi che il sito di legame di un anticorpo per l'antigene è il prodotto di non meno di cinque geni diversi, ciascuno con diverse regioni variabili. I geni si ricombinano quando il linfocito si differenzia e pertanto in ogni cellula vi è una configurazione particolare dalla quale dipende la specificità degli anticorpi che la cellula produrrà.

Abbiamo anche deciso di non occuparci in questa sede della seconda grande famiglia di linfociti, i linfociti *T*. Queste cellule difendono l'organismo, non secernendo anticorpi, ma uccidendo le cellule infette e favorendo il processo infiammatorio. I progressi compiuti nella ricerca sui linfociti *T* sono stati di fondamentale importanza per completare il quadro delle molteplici funzioni del sistema immunitario.

In realtà, il processo di formazione degli anticorpi è appena uno dei numerosi misteri che sono stati risolti. Il sistema immunitario è oggi ben conosciuto a livello cellullare, biochimico e genetico. Vi è ancora molto da imparare, ma i progressi compiuti finora sono di buon auspicio per il futuro. Quando la teoria della selezione clonale per la formazione degli anticorpi fu proposta per la prima volta, sembrava improbabile, quasi incredibile. I concetti rivoluzionari, fatti intravedere da Ehrlich e formalmente sviluppati da Jerne, Talmage e Burnet, non solo stimolarono una notevole massa di ricerche sperimentali, ma permisero anche di intuire aspetti più vasti dell'organizzazione e della funzione delle cellule. La risoluzione del problema dell'origine degli anticorpi ha influenzato lo sviluppo di vaccini e il trapianto di organi, come pure la chimica delle proteine e la biologia molecolare. Sotto questo aspetto, il contributo dei tre scienziati che si sono battuti per la teoria della selezione clonale deve essere misurato non soltanto nel quadro dell'immunologia, ma in quello più vasto della biologia in generale.

La tettonica delle formazioni saline

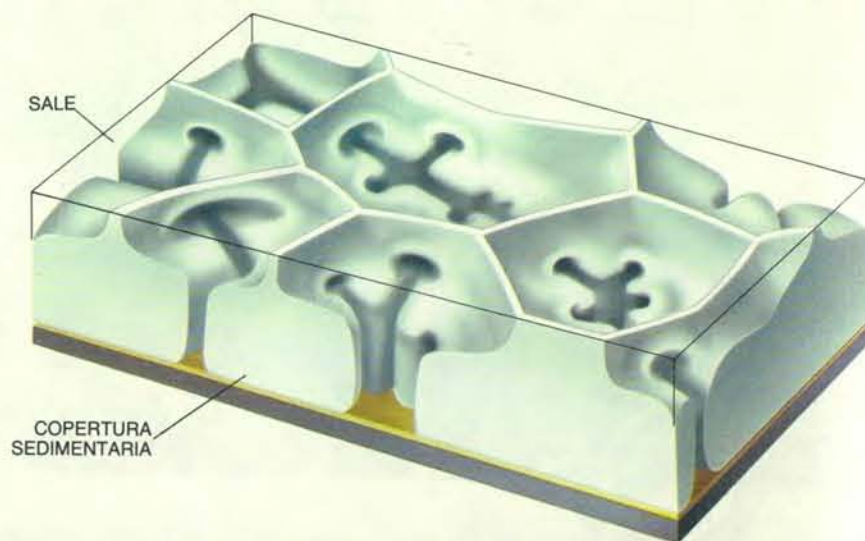
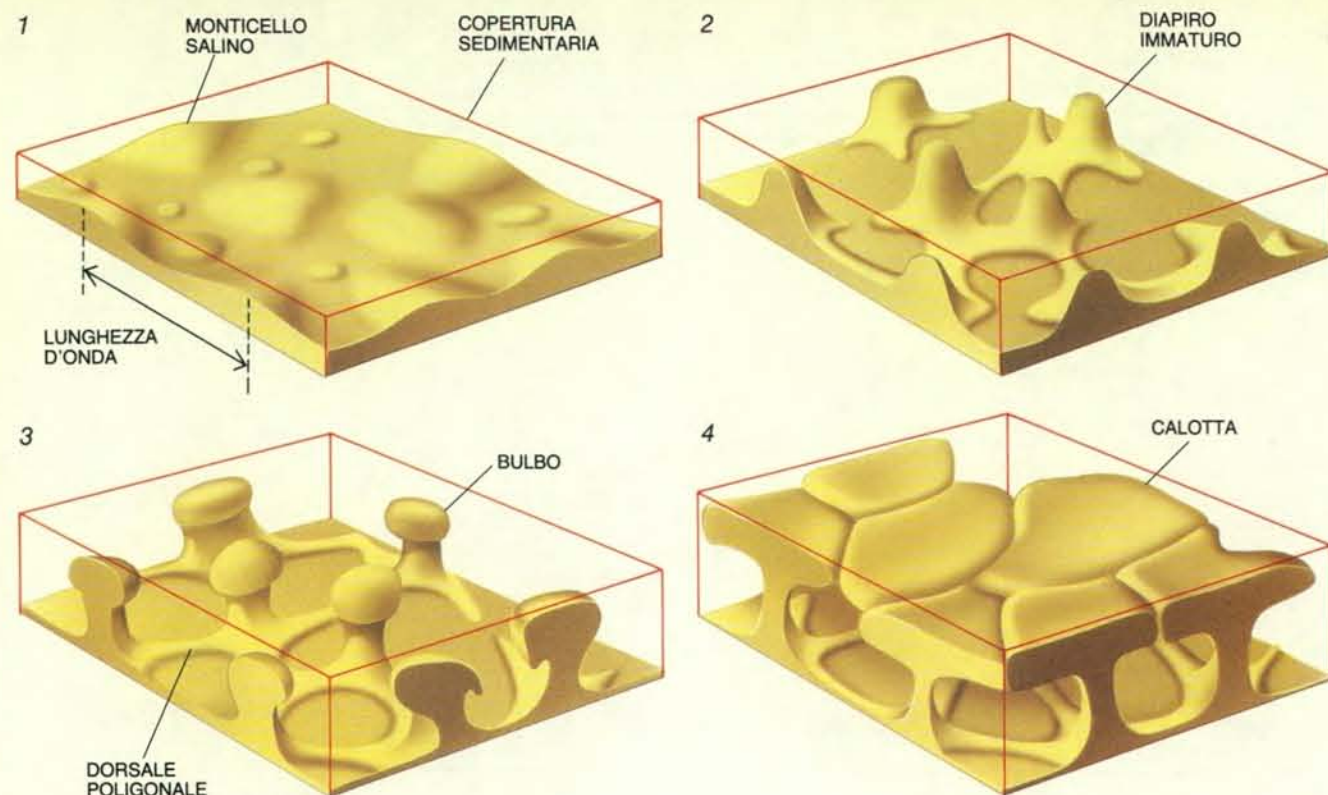
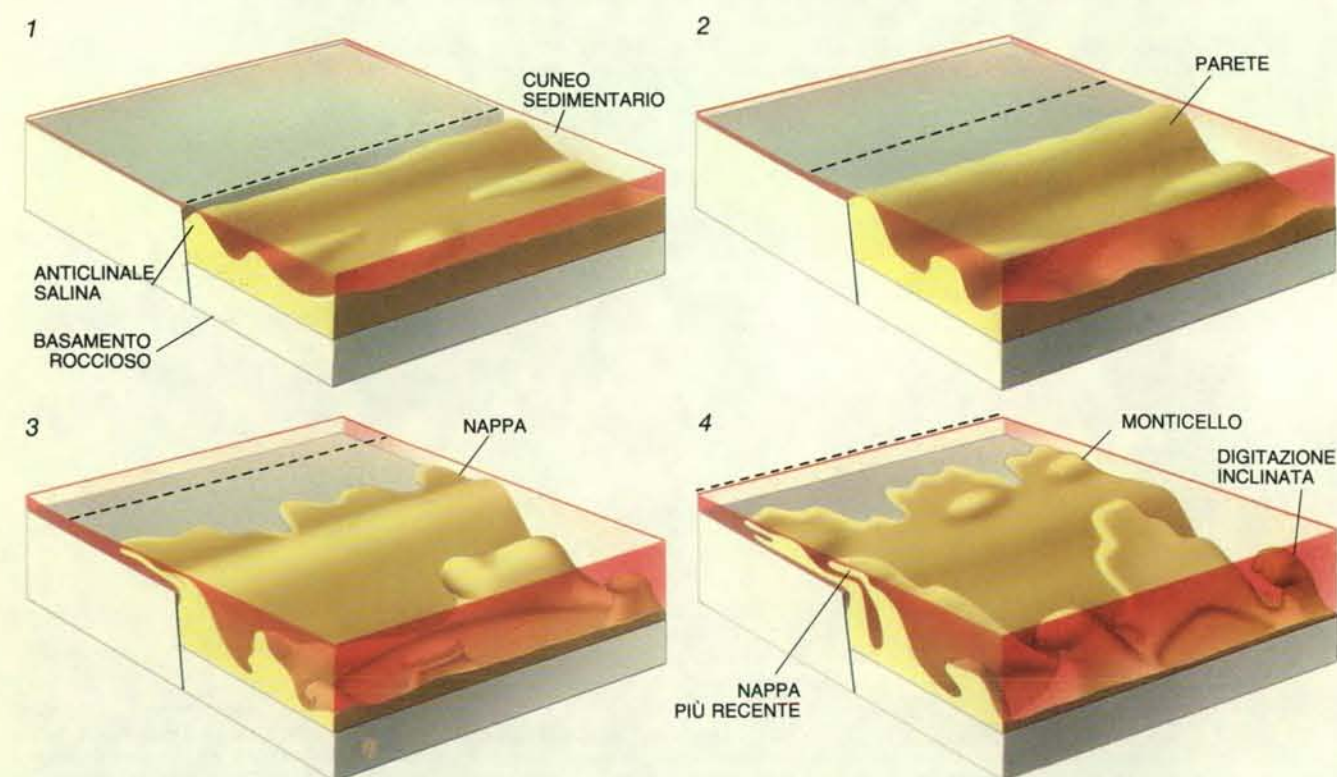
I modelli di laboratorio illustrano come avvenga, in natura, la risalita del sale per spinta idrostatica attraverso formazioni più dense, fino alla fuoriuscita in superficie e allo scorrimento, simile a quello dei ghiacciai

di Christopher J. Talbot e Martin P. A. Jackson

In profondità, al di sotto del Golfo del Messico, enormi digitazioni e protrusioni saline sono in fase di risalita verso il fondo oceanico attraverso 10 chilometri di sedimenti. Presso le coste del Golfo Persico, digitazioni saline provenienti da un deposito più antico sono già fuoriuscite in superficie, dando origine a isolette nel golfo e scorrendo come ghiacciai lungo le pendici dei Monti Zagros. Sollevamenti geologici di tale spettacolarità sono possibili per le proprietà peculiari del sale. La forza che li provoca, per quanto possa sembrare strano, è la gravità.

Il sale è presente in natura in forma di roccia cristallina e sembra quindi logico chiedersi come esso possa scorrere, verso l'alto o altrimenti. È vero che quando viene sottoposto a una sollecitazione improvvisa, come nel caso di una scossa tellurica o di un colpo di martello, il sale reagisce come un solido elastico e può o frantumarsi o rimbalzare, ma quando è soggetto a una sollecitazione continua, anche se di bassa intensità, esso si comporta come un fluido estremamente viscoso e si deforma plasticamente senza rompersi. Si tratta di uno scorrimento molto lento rilevabile solo alla scala geologica dei tempi ed è un fenomeno comune a tutte le rocce.

La maggior parte delle rocce, tuttavia, va incontro a questo lento scorrimento plastico solo per azione delle forze laterali relativamente intense esercitate dal movimento reciproco delle zolle, le stesse forze che provocano il sollevamento delle catene montuose. Il sale, invece, fluisce facilmente anche solo per azione della forza di gravità. Poiché è una delle rocce meno dense ed è quasi incompressibile, il sale rimane leggero anche quando è ricoperto da coltri di sedimenti, quali sabbia, silt e fanghi. I sedimenti sovrastanti, d'altro canto, vengono gradualmente compressi fino a formare rispettivamente dense arenarie, siltiti e argilloscisti. Ne risulta un'inversione di densità: una situazione in cui rocce pesanti stanno al di sopra di altre più leggere è gravitazionalmente instabile. La roccia sedimentaria di copertura inizia allora a sprofondare nel sale, spiazzandolo verso l'alto. Sotto la spinta idrostatica, il sale si fa strada attraverso la copertura formando intrusioni, i cosiddetti diapiri salini.

I diapiri salini sono economicamente importanti. La loro utilità nella prospezione di petrolio e di gas naturale è riconosciuta fin dall'inizio del secolo. Petrolio e gas in risalita sono intrappolati contro i fianchi dei diapiri salini che si sollevano o nei duomi di sedimenti formati dal loro sollevamento. Circa i quattro quinti delle riserve di petrolio e gas degli Stati Uniti meridionali sono associati in qualche modo al sale, come pure alcuni dei grandi campi petroliferi del Medio Oriente. Inoltre greggio e gas già estratti

La struttura a celle della risalita del sale è evidente nell'esperimento condotto da Peter Rönnlund dell'Università di Uppsala. Il modello consiste di due strati: un primo strato denso e colorato di stucco al silicone che rappresenta la roccia sedimentaria di copertura e un secondo meno denso e trasparente di polidimetilsilossano in luogo del deposito salino. Inizialmente gli strati erano orizzontali e quello colorato era alla sommità. Il modello è stato poi fatto ruotare in centrifuga per diversi minuti, per riprodurre (a un tasso accelerato) l'effetto della gravità. La fotografia lo mostra dall'alto. Lo strato colorato si è allontanato dall'osservatore attraverso il «sale» trasparente, che si trova ora al di sopra. Il «sale» è risalito in diapiri sormontati da bulbi a partire dalle triple giunzioni di dorsali poligonali (le fessure scure, ombreggiate della fotografia); i diapiri trasparenti sono separati da avvallamenti poligonali di materiale che sprofonda, i quali appaiono come sottili «pareti» colorate. Le tenui linee in nero, che costituivano una griglia millimetrica sulla faccia inferiore dello strato, indicano la distorsione dello strato di copertura colorato. Il disegno mostra una vista laterale di parte del modello (*in basso a sinistra nella fotografia*).

In questa rappresentazione delle quattro fasi della risalita di un deposito salino, la copertura sedimentaria che si deposita contemporaneamente al processo di risalita è trasparente. Inizialmente l'interfaccia fra i due strati è pressoché piana. Nella prima fase, alcune piccole irregolarità sulla superficie del sale diventano monticelli regolarmente spaziati (di una lunghezza d'onda) (*1*). I monticelli si contraggono in diapiri digitiformi connessi da dorsali saline poligonali (*2*). A piccola profondità i diapiri che si sollevano si espandono in bulbi schiacciati al di sotto di uno strato di sedimenti non compattati e meno densi (*3*). I bulbi possono alla fine unirsi in un'ampia calotta sormontante esili steli salini (*4*).

Gli effetti «di bordo» modificano l'andamento della risalita. L'illustrazione mostra gli effetti combinati di due discontinuità naturali: un gradino di faglia nel basamento roccioso sottostante e il cuneo di sedimenti avanzanti di un delta fluviale. (La linea tratteggiata è l'estremità sottile del cuneo.) Il sale inizia a sollevarsi in un'anticlinale parallela al gradino (*1*); poi si contrae in una parete (*2*). Il cuneo sedimentario avanzante agisce come un matterello, piegando la parete sopra il gradino e formando una nappa salina che cresce di lato (*3*), su cui possono sollevarsi monticelli e crescere nappe più piccole (*4*). Le digitazioni saline in crescita a una certa distanza si inclinano e si rigonfiano verso il gradino.

sono spesso ripompati per immagazzinarli in caverne saline grandi come cattedrali. In futuro anche le scorie radioattive potrebbero essere confinate in formazioni saline. Tutte queste applicazioni forniscono un forte incentivo ad appurare nei dettagli le modalità di evoluzione delle strutture saline.

Per far questo occorre sapere come il sale si forma inizialmente. Esso ha origine nelle zone subtropicali, dove l'aria secca che si è liberata della propria umidità sugli oceani tropicali e sulle foreste pluviali perde quota e si riscalda e dove, di conseguenza, l'evaporazione supera comunemente le precipitazioni. (La maggior parte dei grandi deserti è situata in zone subtropicali.) In bacini sedimentari chiusi, o in mari protetti dall'oceano aperto da scogliere coralline o da altre barriere naturali, l'acqua marina evapora lasciando in loco una salamoia, ossia acqua sovrasalata, dalla quale i minerali precipitano in ragione inversa alla loro solubilità: per primi i carbonati, quindi il gesso (un solfato di calcio), successivamente il salgemma o halite (cloruro di sodio) e infine i sali di magnesio e potassio. Tutti insieme formano le rocce note come evaporiti. In queste il minerale di gran lunga più abbondante è l'halite, o salgemma, che di qui in avanti chiameremo semplicemente «sale». Esso precipita dalla salamoia in cristalli a grana grossolana strettamente uniti.

Se il bacino viene ciclicamente rifornito con acqua marina normale, si possono accumulare sul fondo strati di evaporiti anche di diversi chilometri di spessore. Un massimo nell'accumulo di evaporiti si ebbe circa 230 milioni di anni fa, all'epoca in cui il supercontinente Pangea iniziava a frammentarsi. Fosse tettoniche (*rift valley*) e bacini al di sotto del livello del mare, entro e intorno al Pangea, costituivano ambienti privilegiati per la formazione di evaporiti. Durante quel periodo, qualche cosa come 10 milioni di chilometri cubi di sale, depositatisi da circa 660 milioni di chilometri cubi di acqua marina, formarono strati dello spessore di chilometri. Alcuni dei rift del Pangea si ampliarono in oceani; di conseguenza gli strati di sale sono ora reperibili su alcuni dei margini continentali attuali, sepolti da pile di sedimenti più recenti.

Alla fine, il sale sepolto affiora in superficie ed è riciclato negli oceani dall'erosione. Si sono ritrovati solamente pochi depositi di età superiore a circa 800 milioni di anni e ciò sembra indicare che la maggior parte del sale formatosi prima di allora sia già stata riciclata. Tuttavia il processo è estremamente lento. Il sale che attualmente scende dai Monti Zagros ha un'età compresa fra 500 e 800 milioni di anni, mentre i diapiri che ancora devono raggiungere i fondali del Golfo del Messico provengono da un deposito messo in posto circa 175 milioni di anni fa. In generale, la risalita dei diapiri salini è episodica: periodi attivi, durante i quali i diapiri si sollevano a velocità compresa fra 0,1 e un millimetro all'anno per parecchi milioni di anni (approssimativamente 100 volte più lenti rispetto alla velocità laterale delle zolle crostali), si alternano a periodi di quiescenza.

Questo lento andamento pone i ricercatori di fronte a problemi particolari. Per quanto in laboratorio lo scorrimento del sale possa essere accelerato sottoponendo piccoli campioni a pressioni molto superiori a quelle presenti in natura, questo tipo di esperimento non è in grado di rivelarci né l'andamento del fenomeno di grande scala, né i meccanismi del flusso salino quale avviene in natura. Ciononostante, si può studiare il processo di risalita in vari modi. Un primo metodo consiste nel trovare una soluzione analitica alle equazioni di dinamica dei fluidi che descrivono un sistema generalizzato costituito da un fluido denso (in questo caso la copertura sedimentaria) che ricopre un fluido meno denso (lo strato di origine salina). Anche il secondo metodo è matematico; esso richiede la programmazione di un calcolatore per la risoluzione numerica delle equazioni. Infine - ed è questo il metodo da noi seguito - è possibile simulare la risalita del sale con un modello fisico in scala appropriata.

Il processo di risalita può essere per convenienza diviso in quattro fasi. La prima di esse è ben descritta dalla teoria analitica formulata dai fisici inglesi Lord Rayleigh e Sir Geoffrey Taylor. Secondo la teoria di Rayleigh-Taylor, la risalita ha inizio con impercettibili protuberanze all'interfaccia tra i due fluidi viscosi. Inizialmente le protuberanze crescono a velocità differenti, però solo quelle che crescono più rapidamente sopravvivono perché sottraggono sale alle protuberanze più lente, che vengono pertanto soppresse. A poco a poco, l'interfaccia, inizialmente pressoché piana, si trasforma in un campo di «monticelli» sinusoidali regolarmente spaziati, formati dalle protuberanze che crescono più velocemente. La distanza tra i monticelli (la lunghezza d'onda) è caratteristica del sistema; essa dipende in primo luogo dallo spessore e dalla viscosità relativi dei due strati fluidi. La loro densità relativa determina la velocità di risalita. In natura i monticelli di sale impiegano per formarsi circa 20 milioni di anni.

Durante la prima fase di risalita, gli spazi fra i monticelli sono occupati da avvallamenti, formati dallo sprofondamento della roccia di copertura, che sono in realtà l'immagine speculare dei monticelli. Nella seconda fase le forme dei monticelli e degli avvallamenti iniziano a differire. Appena i monticelli superano una certa altezza, che può variare da circa 0,5 a 2,5 chilometri, si contraggono in strette digitazioni o in «pareti», mentre gli avvallamenti si espandono in

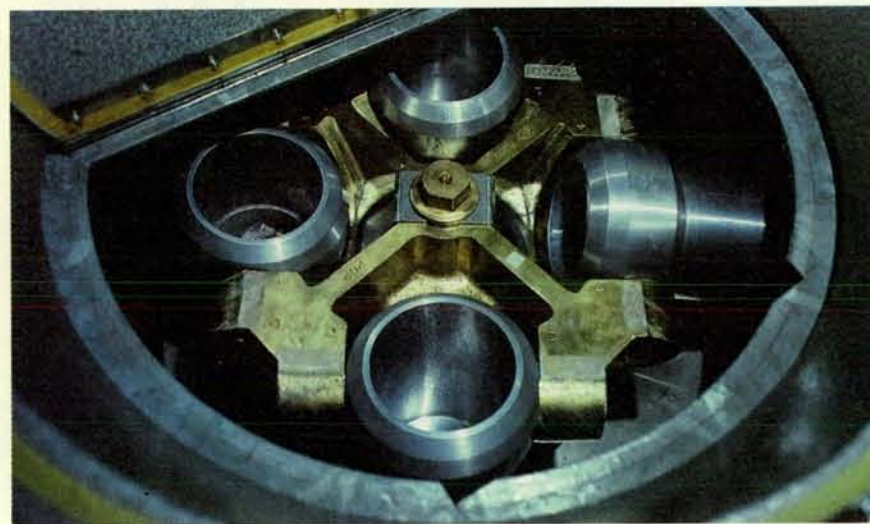
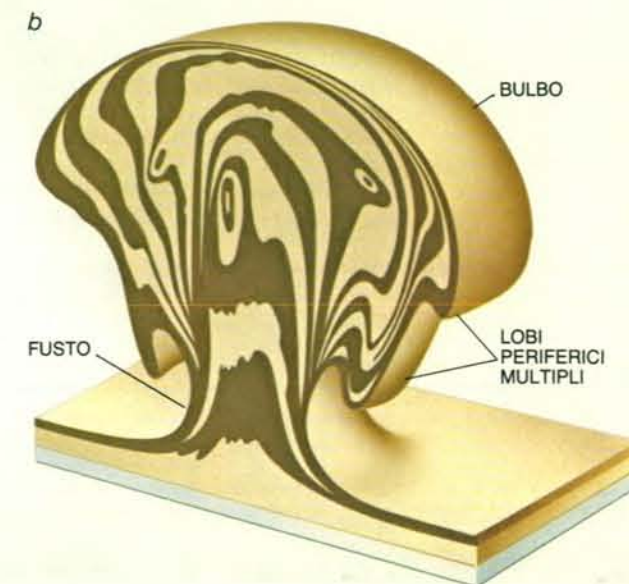
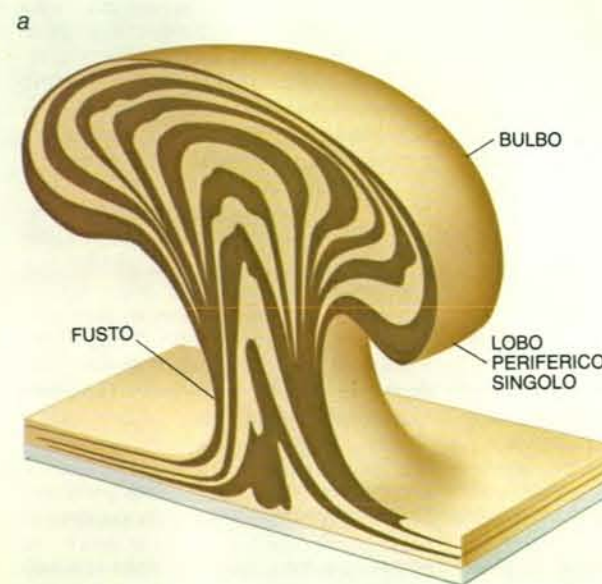
Dalla circolazione associata ai diapiri salini possono avere origine tre tipi di bulbo. Se la copertura sedimentaria è più tenera (meno viscosa) del sale, il bulbo del diapiro assume la forma di un pollice (*a*); se la copertura è più viscosa, il bulbo si conforma a pallone (*b*); se i due strati hanno la stessa viscosità, si ha un bulbo a forma di fungo (*c*). Tutti i tipi di bulbo si possono sviluppare in profondità, prima che il diapiro incontri un margine superiore.

ampi bacini che sprofondano. Alla fine il sale penetra effettivamente nella roccia di copertura in diapiri (così chiamati dal greco *diapeirein* che significa penetrare). In questa fase, nelle equazioni di moto che descrivono la risalita entrano termini non lineari e le equazioni perciò non sono più suscettibili di risoluzione analitica precisa. Per comprendere l'ulteriore evoluzione dei diapiri salini occorre servirsi o di modelli fisici o di modelli al calcolatore.

Nei primi modelli fisici, veniva prodotto un sistema di fluidi gravitazionalmente instabile, ponendo in un recipiente olio galleggiante al di sopra di un denso sciroppo e quindi capovolgendo il recipiente. A causa delle basse viscosità di questi fluidi, nel giro di pochi secondi, si formavano diapiri in risalita facilmente osservabili. D'altro canto era impossibile costruire modelli realistici di rocce sedimentarie dalla complessa stratificazione utilizzando olio e sciroppo. Nel 1960, Hans Ramberg, dell'Università di Upsala, trovò il modo di aggirare il problema. La sua idea fu di costruire modelli con materiali più viscosi, come la creta da modellare e lo stucco al silicone e poi di farli ruotare in una centrifuga.

La forza centrifuga imita l'effetto della gravità, tranne che per il fatto di essere molto più intensa; il materiale meno denso, dapprima al fondo, «risale» all'interno verso l'asse della centrifuga.

Per studiare la risalita del sale abbiamo seguito il metodo di Ramberg. La viscosità e la densità del materiale del modello devono essere tali che un piccolo strato di esso, sottoposto a una forza da 1000 a 2000 volte superiore a quella di gravità per la durata di alcuni minuti, riproduca con precisione le deformazioni subite da un grande deposito salino sottoposto alla normale forza di gravità per milioni di anni. Generalmente usiamo stucco al silicone o polidimetilsilossano in luogo del sale e una miscela di stucco al silicone e di denso solfato di bario in luogo della copertura sedimentaria. Dopo aver fatto ruotare il modello, lo preleviamo dalla centrifuga e ne tagliamo una trentina di sezioni trasversali. Dalle sezioni e da griglie di riferimento fissate nel modello, possiamo ricostruire l'andamento del flusso in tre dimensioni.

Abbiamo trovato che la geometria della risalita di una formazione salina è alquanto complessa; in linea generale è

I modelli sottoposti a centrifugazione rivelano la distribuzione, la forma e la struttura interna dei diapiri salini. Strati di stucco al silicone vengono impilati in una piccola coppa (del diametro di 10 centimetri) e vengono fatti ruotare in una centrifuga (*si veda la fotografia*). Il modello viene quindi rimosso dalla coppa e tagliato in circa 30 sezioni orizzontali o verticali, da cui può essere ricostruito lo schema di flusso tridimensionale. Se tutto il materiale di copertura (che qui non è mostrato) viene aggiunto prima della rotazione in centrifuga del modello, si formano relativamente pochi diapiri (*qui sopra*), e quelli a fungo hanno un unico lobo periferico (*a*). Se gli strati di copertura vengono aggiunti a più riprese per simulare l'accumulazione episodica del sedimento, i diapiri sono molti e si affollano nello spazio disponibile (*in alto a destra*) e i funghi hanno lobi multipli (*b*). Le strisce nelle sezioni del diapiro sono strati di riferimento, inizialmente piani e di uguale spessore. Le pareti ai margini del modello sono originate dagli effetti di bordo.

Il deserto del Gran Kavir, nell'Iran centrale, è sede di diapiri salini spettacolari. La fotografia aerea di Augusto Gansser mostra la sommità troncata di un diapiro quasi circolare, del diametro di sei chilometri. Un nucleo più chiaro di sale incrostato di gesso insolubile appare circondato da evaporiti più recenti sottilmente interstratificate: sale e gesso, di colore chiaro, si alternano con argille e marne più scure. I sedimenti di copertura a stratificazione più grossolana visibili in primo piano sono stati portati verso l'alto dal trascinamento esercitato dai bordi del diapiro.

analoga all'andamento a raggiera caratteristico della convezione termica energica. (Anche la risalita termica è alimentata da un'inversione di densità, ma in questo caso l'inversione è causata da un gradiente di temperatura entro un unico fluido riscaldato dal basso e non dalla differenza di composizione di due fluidi.) Il flusso è ripartito in numerose celle unitarie, ognuna consistente di un solo diapiro e della copertura in sprofondamento al suo intorno. Ogni diapiro si innalza a partire dalla giunzione di tre o cinque dorsali saline (i raggi), che sono connesse in una configurazione tipica a pentagoni o esagoni irregolari. Gli avvallamenti che sprofondano formano a loro volta poligoni centrati sulle creste dei diapiri che si sollevano; ossia, gli avvallamenti sono spostati lateralmente rispetto alle dorsali di una metà di poligono. Il modello poligonale a raggiera non è un mero artificio da noi introdotto: indagini sismiche hanno messo in evidenza la presenza di dorsali poligonali al di sotto della parte nordoccidentale del Golfo del Messico. Ogni poligono misura trasversalmente circa 20-30 chilometri.

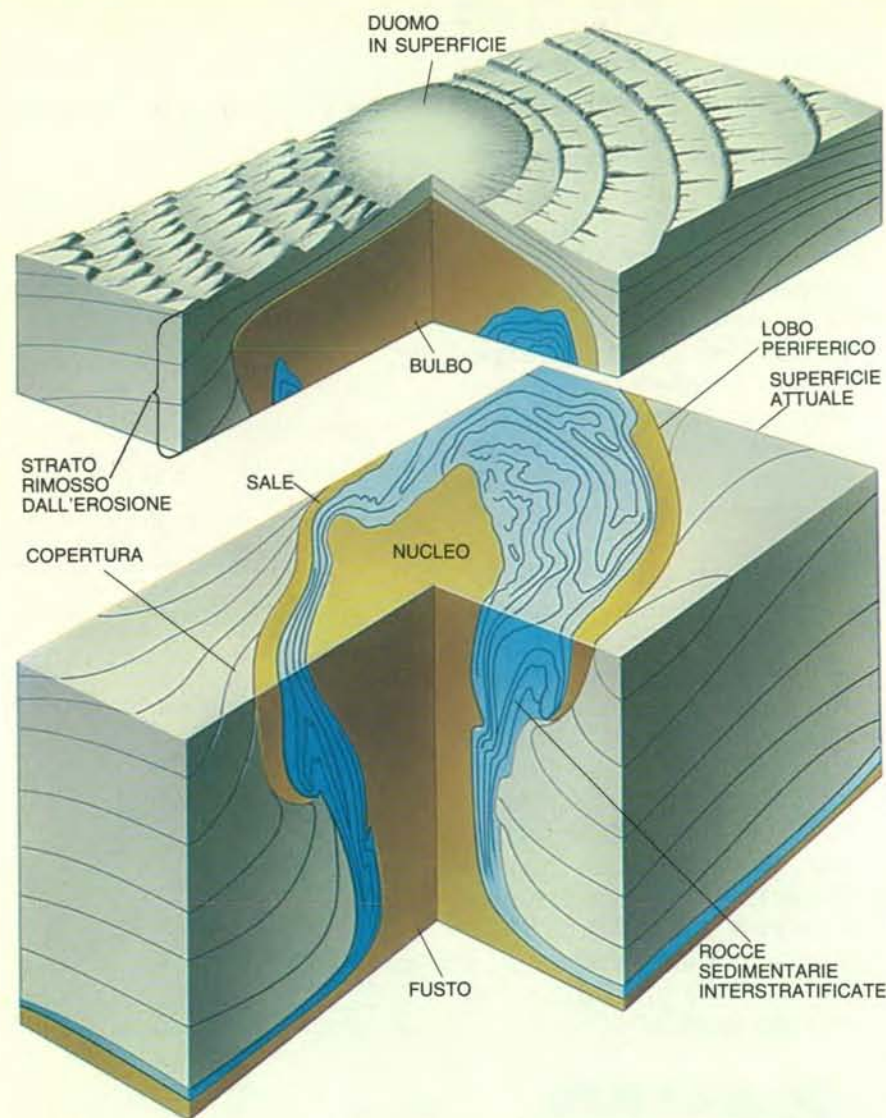
Vicino ai bordi del nostro modello, l'andamento a raggiera non viene rispettato. Anziché risalire in digitazioni, il sale si solleva in lunghe pareti che seguono parallelamente i margini del modello. Nell'Europa nordoccidentale, pareti saline dall'andamento parallelo dimostrano che faglie lineari, pieghe o scarpate nel basamento roccioso sottostante uno strato salino e perfino una discontinuità laterale di carico dovuta a sedimenti fluviali possono dare un effetto naturale di bordo. Ma, lontano dai bordi, siamo convinti che in natura predomini l'andamento a raggiera, così come avviene nei modelli di laboratorio.

In natura però, l'andamento è generalmente modificato da variazioni laterali nello spessore o nelle proprietà meccaniche del sale e degli strati di copertura che sono troppo graduali per poter dare origine a effetti di bordo. Ne risulta che i diapiri salini spesso non sono simmetrici rispetto al loro asse verticale ma inclinati, con un lato che cresce più velocemente ed è più grande dell'altro. L'asimmetricità della crescita può essere esaltata da un effetto di retroazione positiva. Sul lato a rapida crescita di un diapiro fluisce dal deposito sottostante una maggior quantità di sale. Ivi la copertura sedimentaria sprofonderà più rapidamente, dando luogo in superficie a un infossamento che raccoglierà una quantità di sedimenti maggiore; questi, aumentando il carico sul deposito salino, spingeranno una maggiore quantità di sale verso quel lato del diapiro, provocandone una crescita ancora più rapida.

La forma dei diapiri in crescita può essere ancor più irregolare se il deposito salino presenta inizialmente grandi irregolarità - gradini di faglia o ampi rigonfiamenti - alla superficie superiore. Modelli al calcolatore elaborati da Harro Schmeling dell'Università di Uppsala hanno dimostrato che il sale tende a «scaturire» sopra questo tipo di irregolarità, anche se la spaziatura tra di esse è molto più grande o più piccola della lunghezza d'onda tipica del diapiro prevista dalla teoria di Rayleigh-Taylor sulla base dello spessore e della viscosità del sale e degli strati di copertura. Se la spaziatura delle irregolarità è piccola, i diapiri continuano a formarsi a intervalli di una lunghezza d'onda; il sale in risalita al di sopra delle irregolarità, però, viene incorporato nei diapiri, che possono assumere pertanto forme bizzarre.

Il cambiamento di forma più pronunciato cui va incontro un diapiro in sollevamento si ha nella terza fase della sua evoluzione, allorché la sua parte sommitale si dilata in un bulbo sormontante un fusto sottile. I bulbi possono formarsi in due modi. Il primo si ha quando il diapiro incontra un margine superiore, nel caso tipico uno strato sottile di sedimenti che non sia stato ancora compresso a una densità maggiore di quella del sale. Quando un diapiro si solleva in uno strato a bassa densità, la sua sommità si espande, dando luogo a un bulbo che ha generalmente la forma di un pallone sgonfio e schiacciato.

Se la copertura sedimentaria è sufficientemente spessa, si possono formare bulbi salini anche in profondità, che in tal caso non sono dovuti a un'espansione al di sotto di un margine superiore, ma all'interazione tra il nucleo del diapiro che sale e la copertura che, sprofondando, esercita contro i lati del diapiro un'azione di trascinamento verso il basso. Questa interazione instaura nel diapiro una circolazione interna.

La natura della circolazione e la forma del bulbo che ne risulta dipendono principalmente dalle viscosità relative del sale e della copertura. Fondamentalmente si possono avere tre forme (*si veda l'illustrazione a pagina 57*). Se la copertura è meno viscosa del sale, il diapiro si spinge con facilità attraverso di essa e si stabilisce solo una debole circolazione interna; il diapiro assume una forma a pollice, con un bulbo poco pronunciato. Viceversa, se la copertura è più viscosa e relativamente indeformabile, il sale in risalita viene forzato in un deciso flusso toroidale, che sostiene un bulbo a forma di pallone. Infine, se la copertura e il sale hanno approssimativamente la stessa viscosità, si stabilisce anche in questo caso una circolazione toroidale, che questa volta però interessa entrambi gli strati. La copertura che sprofonda trascina la parte periferica del diapiro in una sorta di lobo pendulo, che risale meno vigoro-

La struttura interna di un diapiro a fungo del Gran Kavir è stata esposta dall'erosione. Mediante estrapolazione delle caratteristiche superficiali e sulla base dei modelli di laboratorio, gli autori hanno dedotto sia la struttura profonda del diapiro, sia quella della sommità asportata dall'erosione. Nel bulbo, la sequenza di stratificazione del deposito originario è invertita: il sale giace sopra rocce sedimentarie interstratificate, che sono intermedie, per età e composizione, tra il sale stesso e la copertura più recente. Tali rocce formano un anello alla superficie attuale.

samente del nucleo; in questo caso il diapiro assume una conformazione a fungo.

In miniere di sale degli Stati Uniti e dell'Europa si sono raccolti dati sulla circolazione interna in diapiri a fungo. Carte e sezioni verticali mostrano come la sequenza di stratificazione del sale nei depositi originali sia stata invertita e rovesciata nei diapiri dal flusso toroidale interno. Le sezioni verticali suggeriscono inoltre che i lobi periferici dei «funghi» possano aver trascinato con sé le rocce di copertura, piegandole in profondità nei diapiri. I nostri modelli indicano come sia probabile che il trascinamento abbia luogo quando molti diapiri dalle forme fortemente irregolari si affollano in una stessa zona. In questa situazione, i diapiri contigui interagiscono con la stessa regione del materiale di copertura che sprofonda e il trascinamento verso il basso esercitato sui diapiri viene così intensificato.

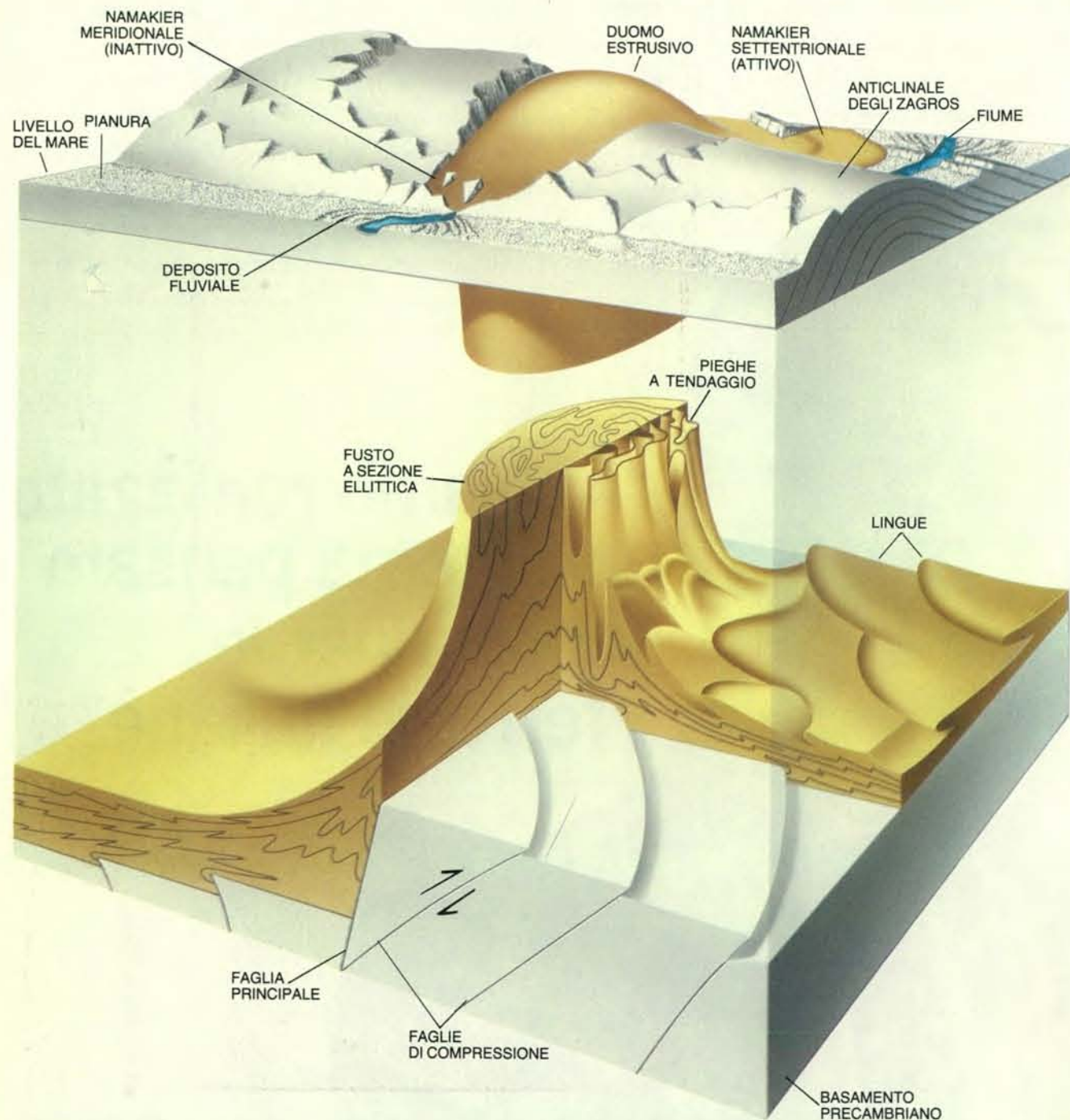
L'affollamento, a sua volta, si crea con tutta probabilità quando la deposizione della copertura sedimentaria avviene episodicamente. Le ragioni sono semplici. Secondo la teoria di Rayleigh-Taylor, la distanza fra i diapiri in un dato deposito salino è proporzionale (a parità di altri fattori) allo spessore della copertura. Se la copertura ha una deposizione ritmica, tuttavia, i diapiri salini potrebbero svilupparsi e giungere a maturità durante i primi episodi di sedimentazione, quando la copertura è ancora relativamente sottile. Il campo di diapiri risultante sarà molto più «affollato» di quanto sarebbe stato se la copertura si fosse depositata in sequenza continua. Se la copertura ha all'incirca la stessa viscosità del sale, i diapiri saranno a forma di fungo e implicheranno probabilmente rocce di copertura sotto i loro lobi.

Questa conclusione ha un significato pratico. Una delle ragioni per le quali le formazioni saline sono state proposte come siti di confinamento per le scorie radioattive è che esse sono relativamente impermeabili. Eventuali fessurazioni vengono sigillate dallo scorrimento plastico e, per quanto il sale sia altamente solubile, l'acqua freatica non potrebbe penetrare molto all'interno senza saturarsi e non essere più in grado di sciogliere altro sale. Se, tuttavia, una caverna per il confinamento di scorie fosse costruita in un diapiro a fungo gli strati incorporati di roccia sedimentaria permeabile potrebbero fungere da condotti, convogliando l'acqua attraverso la caverna. L'acqua potrebbe potenzialmente trasportare radionuclidi nell'ambiente. È chiaro che la forma e la struttura interna di un diapiro salino dovranno essere attentamente considerate prima di sceglierlo come sito per il confinamento di scorie radioattive.

La quarta e ultima fase della risalita di sale ha inizio quando il diapiro a forma di bulbo, sollevandosi, si avvicina alla superficie. A quel punto si avranno tre possibilità. La prima è che i diapiri vengano «decapitati» dall'acqua freatica. Gli strati prossimi alla superficie sono generalmente porosi e l'acqua freatica non satura, in essi contenuta, è in grado di dissolvere la sommità di un diapiro alla stessa velocità con cui esso si solleva. In casi simili la superficie di dissoluzione è ricoperta da un residuo di gesso relativamente insolubile che si trovava disperso nel sale in risalita.

Una seconda possibilità si ha nei deserti, dove la dissoluzione del sale da parte dell'acqua freatica avviene ovviamente con minore probabilità: i bulbi possono continuare a espandersi orizzontalmente al di sotto della barriera costituita da strati superficiali di bassa densità. Nel deserto del Gran Kavir, nell'Iran centrale, i bulbi di 12 diapiri si sono fusi in un'unica calotta di 40 chilometri di diametro. La calotta è stata esposta in superficie dall'erosione ed è osservabile in maniera straordinariamente dettagliata. Essa rappresenta una fase insolitamente avanzata di risalita, in cui la quasi totalità di un deposito salino dell'età di 50 milioni di anni si è scambiata di posto con i sedimenti più densi che inizialmente lo sovrastavano. Sebbene i liquidi viscosi si rovescino completamente se si trovano in uno stato gravitazionalmente instabile, essi sono differenti dalle rocce, in quanto non offrono resistenza allo snervamento e fluiscono pertanto sotto qualsiasi forza agisca su di essi. Sembra che la maggior parte dei diapiri «grippi» prima di fondersi in una calotta, poiché

le forze idrostatiche diventano troppo deboli. A ogni modo, certi strati salini, che ora si suppone siano giovani e indeformati, potrebbero essere in realtà calotte diapiriche di più antica data.

Dove gli strati poco profondi sono più densi del sale, i diapiri possono andare incontro a un terzo destino, riuscendo a farsi strada fino in superficie ed erompendo in lente, maestose fontane. I migliori esempi di estrusioni saline in corso sono i diapiri dei Monti Zagros, nell'Iran meridionale. Durante gli ultimi 15 milioni di anni la risalita è stata qui accelerata dalla tettonica delle zolle: la collisione fra la zolla arabica e la zolla eurasiatica, iniziata circa 15 milioni di anni fa, sta «spremendo» il sale verso l'alto. Circa 20 diapiri sono stati spinti al di sopra della superficie del Golfo Persico, e ora formano piccole isole saline. Sui Monti Zagros, crateri ostruiti da detriti rocciosi insolubili sono vestigia di diapiri che non sono stati in grado di competere con la pioggia. Lungo la costa, dove la collisione fra le zolle è generalmente di massima intensità, montagne a pieghe sono perforate da fontane di sale alte anche più di 1000 metri. Il sale, spinto dal suo stesso peso, scorre lungo le pendici delle montagne formando «colate» chiamate *namakier* (da *namak*, che in lingua parsik sta per «sale» e *glacier*).

I namakier dei Monti Zagros, nell'Iran meridionale, ebbero origine da una fontana salina risalita attraverso sette chilometri di rocce di copertura. Il sale estruso fluisce oggi lungo le pendici del Kuh-e-Namak (28,2 gradi di latitudine nord, 51,7 gradi di longitudine est) alla velocità di pochi metri all'anno, sovrapponendosi a depositi fluviali recenti. Il diapiro che alimenta i namakier si alza al di sopra di una faglia principale del basamento, che ha creato un gradino alla superficie superiore del deposito salino. Il sale che fluisce nel diapiro passa sopra una serie di faglie di compressione e viene probabilmente piegato in lingue sovrapposte; esse sono ruotate lungo la verticale e ripiegate in forme simili a tendaggi.

La velocità media di un namakier, di qualche metro all'anno, è minore di quella della maggior parte dei ghiacciai, ma è nondimeno notevolmente elevata per una roccia cristallina. Come si spiega la rapidità del flusso? Secondo alcuni ricercatori che si basano su esperimenti effettuati con sale privo di umidità, il sale di un namakier verrebbe estruso incandescente. L'ipotesi non sembra ora plausibile. Chi la sosteneva trascurava il fatto, noto da tempo, che se si aumenta il contenuto in acqua del sale se ne riduce la viscosità e la resistenza al flusso, altrettanto efficacemente di quanto si ottiene con un incremento di temperatura. Le nostre misurazioni su riferimenti disposti su un namakier hanno dimostrato che esso rimane stazionario per la maggior parte dell'anno; dopo essere stato impregnato da scrosci stagionali, può però fluire in ragione di mezzo metro al giorno.

Osservando la deformazione di evaporiti al microscopio, Janos L. Urai e Christopher J. Spiers, dell'Università di Utrecht, hanno studiato l'effetto dell'acqua sullo scorrimento del sale. Sia che l'acqua provenga dalla salamoia originaria da cui si è formato il sale, dai sedimenti circostanti o da precipitazioni piovose, essa tende a formare una sottile pellicola continua lungo i bordi dei grani quando il sale viene deformato. La pellicola indebolisce il sale drasticamente. In particolare, essa passa attraverso i grani vecchi e distorti, le cui strutture interne dislocate resistono a ulteriori deformazioni, e li dissolve; gli ioni sodio e cloro diffondono attraverso la pellicola acquosa e danno luogo a nuovi grani facilmente deformabili. Un basso contenuto in acqua, anche dello 0,1 per cento in peso, è sufficiente per favorire questa «ricristallizzazione dinamica», la qual cosa suggerisce che il processo non avvenga solo nei namakier, ma anche in profondità, in depositi salini sepolti.

Il sale che fluisce a valle in un namakier è stato soggetto a una lunga e complessa storia di ripiegamenti e di altre deformazioni. Le tracce di questa storia sono registrate nella stratificazione a bande colorate dovute alla deposizione, insieme all'halite, di concentrazioni variabili di altri minerali. La più recente direzione di flusso è rivelata dall'allineamento dei grani di minerale. Nel namakier che abbiamo studiato in maniera approfondita l'allineamento dei grani mostra che la direzione del flusso è in genere parallela sia alla stratificazione interna sia alle pareti del rigido alveo che contiene il namakier.

Scendendo lungo la montagna, però, il sale passa sopra una serie di irregolarità a gradini del fondo dell'alveo. Prima di ogni gradino, il flusso rallenta e si ispessisce, le linee di flusso divergono e gli strati colorati vengono distorti in una serie di pieghe. Dopo l'ostacolo, il flusso accelera nuovamente e le linee di flusso convergono. Gli strati si assottigliano di

La costa del Golfo Persico parallela alle pieghe degli Zagros è punteggiata di creste diapiriche. La fotografia è composta di due immagini elaborate con dati Landsat; la regione mostrata in figura si trova circa 100 chilometri a ovest dello stretto di Hormuz e ha una larghezza di circa 150 chilometri. Le pieghe parallele alla linea di costa sono i Monti Zagros. I diapiri salini assomigliano a scure teste di cavolfiore. La loro risalita in superficie è stata accelerata dallo stesso processo che ha innalzato le montagne: la collisione in corso fra le zolle arabica ed eurasiatica. Tre delle isole, che prima o poi saranno vette sulla terraferma, hanno un nucleo salino.

nuovo, le pieghe interne subiscono uno stiramento e le laminazioni saline cominciano a scorrere l'una sull'altra. Al gradino seguente, nuove pieghe si sovrimpongono alle vecchie. Da qui in poi, il namakier consisterà in una pila di lingue sovrapposte. Nella punta del namakier la pila è così appiattita che le pieghe sono poco visibili; le lingue appaiono come strati orizzontali pressoché indeformati.

Effetti simili hanno probabilmente luogo anche al di sotto della superficie, nel deposito salino profondo che alimenta i diapiri. Ovunque il sale che migra verso un diapiro incontri irregolarità nel tetto o nel letto del deposito, si possono generare lingue interne dalla giacitura piana. Lingue provenienti da un'ampia area di deposito sono convogliate nel sottile fusto del diapiro, dove ognuna di esse viene ruotata lungo la verticale e costretta in pieghe a tendaggio. (L'effetto è simile a quello che si ottiene facendo passare un tovagliolo attraverso un portatovagliolo ad anello.) Il piegarsi e ripiegarsi di parecchie generazioni di lingue spiega l'ampia varietà di forme splendide e intricate che si possono osservare sulle volte delle miniere di sale.

Per chi abbia studiato per un po' le formazioni saline, è difficile esimersi da un'ammirazione puramente estetica: la comprensione della strana anatomia di queste strutture quasi organiche è già di per sé una ricompensa al lavoro. Ma, come già in precedenza abbiamo sottolineato, questi studi hanno anche valore economico. La possibilità che alcuni diapiri abbiano una conformazione a fungo ha un grande significato per la ricerca di petrolio e di gas naturale perché molti diapiri potrebbero essere delle «trappole» per il petrolio considerevolmente più grandi di quanto si sia mai pensato. Esattamente allo stesso modo in cui i lobi di un diapiro a fungo potrebbero includere rocce permeabili che comprometterebbero la tenuta di una caverna per il confinamento delle scorie radioattive, così potrebbero includere o celare rocce serbatoio impregnate di petrolio.

Il fatto che i diapiri a fungo siano stati scoperti in laboratorio illustra l'importanza della costruzione di modelli. Gli scavi e le perforazioni possono fornire solo un numero limitato di informazioni sulle strutture saline sotterranee, e, in generale, anche i dati di campagna disponibili non sono stati sfruttati appieno, semplicemente perché i ricercatori non sapevano come interpretarli. Le cose stanno attualmente cambiando, in quanto i modelli fisici e quelli al calcolatore stanno mettendo in luce i processi fondamentali che provocano la risalita del sale. In futuro, tanto i geologi che si occupano di ricerche petrolifere quanto i tecnici che studiano i siti per il confinamento delle scorie radioattive trarranno vantaggio da una comprensione molto più approfondita delle strutture geologiche che intendono sfruttare.

I transistori di arseniuro di gallio

Indispensabili per sistemi di calcolo e comunicazione ad altissima velocità, hanno caratteristiche che possono essere spiegate studiando l'energia e la quantità di moto di un elettrone che si propaghi nel cristallo

di William R. Frensley

Nei dispositivi a semiconduttore la velocità è di enorme importanza. Nei calcolatori attuali, i dispositivi più veloci possono commutare, cioè aprire e chiudere, la corrente in un miliardesimo di secondo circa. Tuttavia dispositivi ancora più veloci sono essenziali per costruire calcolatori più potenti e consentirebbero inoltre la costruzione di radar e di satelliti per comunicazioni funzionanti nella gamma delle microonde e a frequenze ancora più elevate.

Per soddisfare queste esigenze, nei laboratori di tutto il mondo si stanno sperimentando nuove tecnologie per i semiconduttori. Un metodo per accrescere la velocità dei dispositivi a semiconduttore è di renderli minuscoli, diminuendo così la distanza che gli elettroni che portano i segnali devono percorrere. Un secondo metodo consiste nell'aumentare la velocità degli elettroni che vi si propagano. Le tecnologie che adottano questa impostazione in genere non si basano sull'elemento semiconduttore tradizionale, il silicio, bensì su un composto semiconduttore, l'arseniuro di gallio.

I progettisti dei dispositivi ad arseniuro di gallio non possono basarsi soltanto sul vasto complesso di conoscenze teoriche e pratiche su cui si basa l'affermata tecnologia del silicio. In un cristallo di arseniuro di gallio, le cariche dei nuclei atomici e la distribuzione degli elettroni differiscono notevolmente da quelle degli atomi di un cristallo di silicio. I fisici che studiano i semiconduttori hanno perciò sviluppato metodi che rappresentino gli effetti che queste differenze hanno sulle proprietà ondulatorie di un elettrone che attraversi un cristallo di arseniuro di gallio. Queste rappresentazioni si basano sulle soluzioni numeriche delle equazioni quantomeccaniche che governano la dinamica degli elettroni.

Grazie a queste e ad altre tecniche, è stato possibile comprendere e sfruttare le singolarissime proprietà elettroniche dell'arseniuro di gallio. La più importante di queste proprietà è il fatto che un elettrone che si propaghi in un cristallo di arseniuro di gallio si comporta come se avesse una massa minore di quella che sembra possedere in un cristallo di silicio. Quindi un campo elettrico di data intensità accelera di più gli elettroni nell'arseniuro di gallio che nel silicio. Tale proprietà, e altri effetti quantomeccanici più bizzarri osservati nell'arseniuro di gallio, possono essere sfruttati per costruire dispositivi a semiconduttore aventi, sotto il profilo della velocità, prestazioni che i dispositivi tradizionali a base di silicio non possono uguagliare.

Per apprezzare i vantaggi dell'arseniuro di gallio, è utile sapere come vengono impiegati i materiali semiconduttori per costruire dispositivi elettronici allo stato solido. Il dispositivo più noto e di più vasto impiego è il transistore, nome col quale in realtà vengono indicate parecchie ampie categorie di dispositivi. Nella sua forma più elementare un transistore è un interruttore elettronico: una corrente che lo attraversa viene avviata o arrestata applicando o togliendo un'altra corrente, più debole, in una zona particolare del transistore. È anche possibile regolare l'intensità della corrente con maggior precisione, in modo che le variazioni della debole corrente applicata inducano variazioni corrispondenti nella corrente che attraversa il transistore. In tal caso quest'ultimo funge da amplificatore elettronico: la corrente più forte che attraversa il transistore varia seguendo l'andamento della più debole corrente applicata.

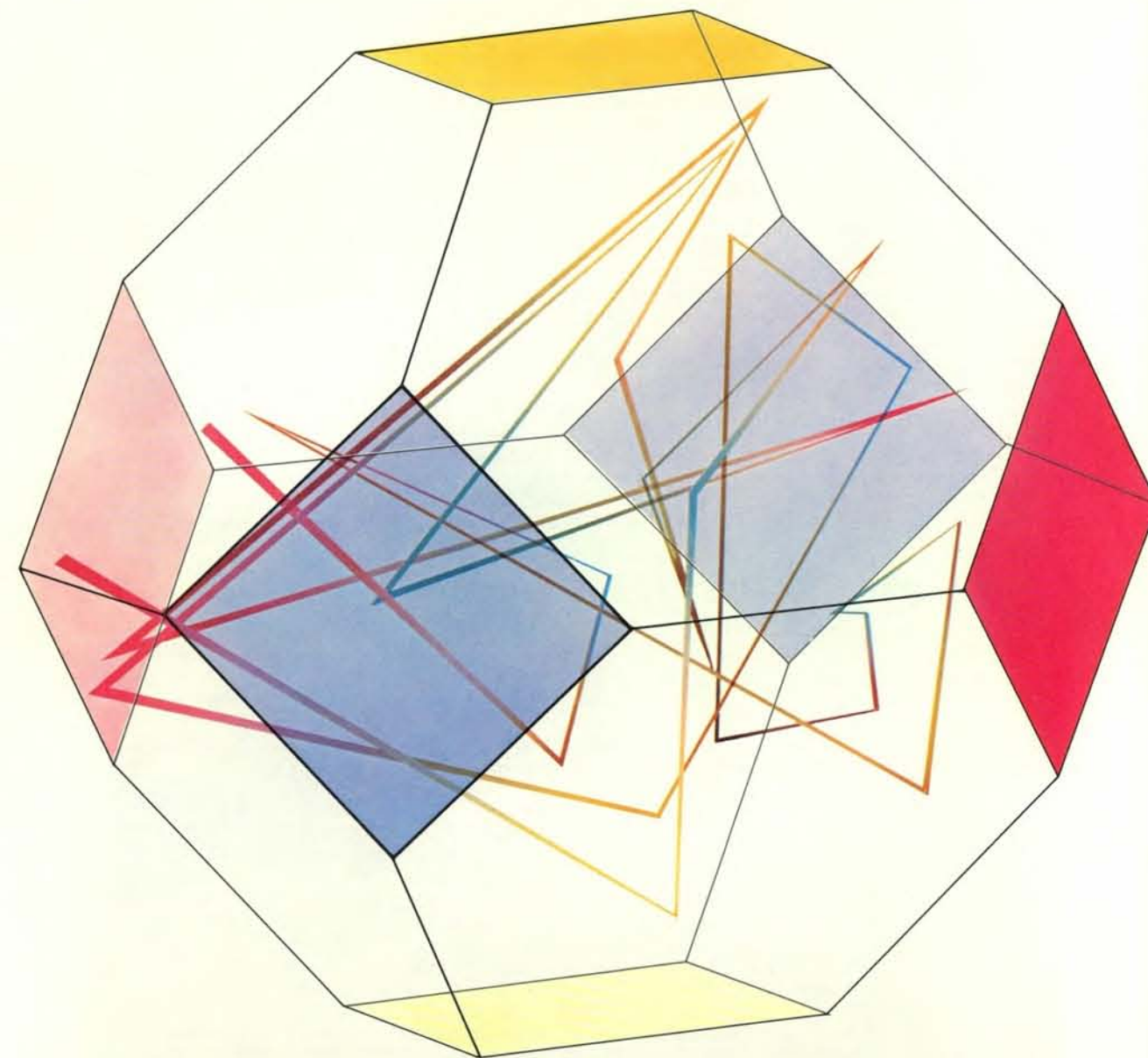
Uno dei dispositivi più semplici di questo genere, il transistore a effetto di campo o FET (da *field-effect transistor*), può servire a illustrare i fattori principali che determinano la velocità di un transistore. Un FET è composto di due strati, ricavati dallo stesso cristallo semiconduttore: uno strato non conduttore collocato sotto uno strato conduttore, detto strato attivo, e che ha uno spessore di 0,1 o 0,2 micrometri. Lo strato attivo viene di solito reso conduttore introducendovi un numero relativamente piccolo di atomi di impurezze, che fungono da donatori di elettroni, cioè cedono un elettrone e si caricano positivamente.

Sulla superficie dello strato attivo vi sono tre elettrodi, chiamati sorgente, pozzo e porta. Una differenza di potenziale applicata fra sorgente e pozzo crea nello strato attivo un campo elettrico e fa muovere gli elettroni lungo le linee del campo. Di solito, la sorgente è a potenziale più basso rispetto al pozzo, sicché gli elettroni vengono iniettati nello strato attivo dalla sorgente e, dopo averlo attraversato, vengono raccolti dal pozzo.

La porta è fatta in modo che la giunzione fra il metallo dell'elettrodo e il semiconduttore sottostante formi quella che è conosciuta come barriera di Schottky. La caratteristica fondamentale di una barriera di Schottky è che gli elettroni del metallo hanno un livello energetico molto minore di quello degli elettroni del semiconduttore adiacente. Di conseguenza, tutti gli elettroni che dal semiconduttore passano nel metallo vi restano intrappolati. Gli elettroni imprigionati nella porta, tuttavia, sono ugualmente attratti verso i donatori positivi del semiconduttore sottostante e restano vicini alla giunzione (a pochi decimillesimi di micrometro), formando sull'elettrodo una carica superficiale negativa. Via via che si forma, questa carica negativa respinge gli elettroni dello strato attivo e, sotto la porta, si forma uno strato di svuotamento, cioè una regione dello strato attivo dalla quale gli elettroni di conduzione sono stati allontanati.

La profondità dello strato di svuotamento viene controllata dalla tensione applicata alla porta. Se questa tensione è resa negativa, ossia se il circuito esterno fornisce alla porta un numero maggiore di elettroni, lo strato di svuotamento si addentra più profondamente nello strato attivo. Quando la tensione diventa sufficientemente negativa, lo strato di svuotamento si estende a tutto lo strato attivo: in tal caso tra sorgente e pozzo non può esserci passaggio di corrente e si dice che il transistore è stato «interdetto». Se la tensione della porta diviene meno negativa, lo strato di svuotamento si assottiglia e tra sorgente e pozzo si ha il passaggio di una corrente più intensa.

Come viene definita la velocità di funzionamento di questo tipo di transistore? Dipende dal circuito al quale esso appartiene. In un circuito logico digitale, come se ne trovano nei calcolatori e nelle calcolatrici, la velocità di un FET è definita come il tempo necessario per interdire il transistore. Si può dimostrare facilmente che ciò equivale al tempo che un elettrone dello strato attivo impiega in media per percorrere la lunghezza della porta.

La velocità di un circuito analogico, come quelli che si trovano nei ricevitori e nei trasmettitori radio e televisivi, è definita in modo alquanto diverso. In un circuito di questo genere la funzione del transistore non consiste nel permettere o meno che la corrente giunga al pozzo, bensì nel trasformare le piccole fluttuazioni della corrente della porta in fluttuazioni corrispondenti e più ampie della corrente che va al pozzo. Il parametro fondamentale è il guadagno di corrente del transistore, o amplificazione, che è definito come il rapporto fra la corrente del pozzo e la corrente della porta. Il guadagno di corrente diminuisce all'au-

La traiettoria di un elettrone in moto all'interno di un cristallo di arseniuro di gallio, sotto l'influsso di un campo elettrico intensissimo (*linee in colore*), può essere tracciata in funzione della quantità di moto dell'elettrone secondo tre direzioni ortogonali. In questo «spazio delle quantità di moto», l'elettrone è confinato entro un poliedro, la zona di Brillouin, che contiene tutti i valori possibili della quantità di moto. La forma della zona di Brillouin dipende dalla struttura del reticolo cristallino; la sua regolarità geometrica rispecchia la regolarità del reticolo. Il colore di un punto particolare della traiettoria dell'elettrone indica la somma delle componenti della sua quantità di moto secondo ciascuno dei tre assi, colorati diversamente. Poiché l'elettrone viene diffuso dagli atomi di impurezze e dai fononi (quanti di vibrazione termica), il suo percorso appare casuale. Però, se le componenti della quantità di moto dell'elettrone venissero sommate su un intervallo di tempo abbastanza lungo, si troverebbe che, in effetti, esso si sposta principalmente secondo la direzione del campo elettrico. L'illustrazione si basa su simulazioni al calcolatore effettuate da Hisashi Shichijo della Texas Instruments, Inc., e da Karl Hess dell'Università dell'Illinois a Urbana-Champaign.

mentare della frequenza delle fluttuazioni della corrente della porta. In questo caso la velocità è definita come la capacità di mantenere un guadagno elevato ad alte frequenze. In sostanza, però, sia nei circuiti digitali sia in quelli analogici si dice che il transistore è veloce quando risponde rapidamente a una variazione della corrente della porta, e il metodo principe per aumentarne la velocità in entrambe le applicazioni consiste nel diminuire la lunghezza della porta o nell'aumentare la velocità degli elettroni.

Nei circuiti integrati odierni, un FET a base di silicio ha un elettrodo porta lungo circa un micrometro. Non c'è dubbio che i progressi nelle tecnologie di fabbricazione dei semiconduttori consentiranno un'ulteriore riduzione delle dimensioni della porta e quindi un aumento di velocità dei transistori. Ma, per conseguire le più elevate velocità di funzionamento possibili, è necessario accrescere anche la mobilità degli elettroni all'interno del transistore. Ciò si può ottenere fabbricando circuiti integrati con materiali semiconduttori diversi dal silicio.

Per comprendere quali fattori influiscano sulla velocità degli elettroni in un dato semiconduttore, si deve considerare come essi si muovono entro la struttura cristallina. Quando più atomi vengono avvicinati per formare un cristallo di semiconduttore, i livelli energetici discreti degli elettroni che circondano il nucleo di ogni atomo si allargano e formano fasce energetiche, chiamate bande. I livelli energetici che erano occupati dagli elettroni più esterni degli atomi, gli elettroni di valenza, divengono le cosiddette bande di valenza del cristallo.

Gli elettroni nelle bande di valenza di un semiconduttore, che sono quasi del tutto occupate, formano il reticolo di legami chimici che assicura la coesione del cristallo. Un intervallo di «energia proibita», o banda proibita, in cui non esistono stati elettronici consentiti, separa le bande di valenza dalle bande energetiche non occupate, chiamate bande di conduzione. Poiché gli elettroni contenuti nelle bande di conduzione non partecipano ai legami interatomici, tutti gli elettroni che si trovano in una banda di conduzione sono liberi di muoversi entro il cristallo.

Nell'ambito della meccanica quantistica, una particella, per esempio un elettrone, può essere concepita anche come un'onda. In particolare, un elettrone in moto all'interno di un cristallo può essere trattato come un'onda che si propaga, e la sua lunghezza d'onda decresce al crescere della sua quantità di moto. L'energia totale di un elettrone in un cristallo è data dalla combinazione della sua energia cinetica con l'energia potenziale risultante dalle sue interazioni con tutte le altre particelle cariche che lo circondano. Quindi lo stato energetico di un elettrone dipende sia dalla sua funzione d'onda sia dalla struttura periodica del reticolo cristallino, formata dai nuclei positivi degli atomi che lo costituiscono e dai loro elettroni di valenza.

Un transistore a effetto di campo viene fabbricato con due strati di materiale semiconduttore: lo strato inferiore è semisolante, mentre quello superiore, detto strato attivo, è conduttore. Sulla superficie dello strato attivo vengono formati tre elettrodi metallici. Gli elettroni possono muoversi facilmente nello strato attivo fra i due elettrodi esterni, detti sorgente e pozzo. L'elettrodo intermedio, detto porta, è fatto in modo da respingere gli elettroni del semiconduttore sottostante, formando una zona (lo strato di svuotamento) priva di elettroni di conduzione. Se fra sorgente e pozzo viene applicata una tensione (*in alto*), gli elettroni provenienti dalla sorgente vanno al pozzo attraversando lo strato attivo. (Le linee in colore rappresentano il gradiente di potenziale per incrementi di mezzo volt e le frecce indicano il verso e l'intensità della corrente.) Il flusso di corrente che attraversa il transistore può essere regolato variando la tensione applicata alla porta, poiché una tensione di porta negativa fa scendere più in profondità nello strato attivo lo strato di svuotamento. Se la tensione della porta è sufficientemente negativa, il passaggio di corrente si interrompe (*in basso*). La densità degli elettroni di conduzione è indicata dalla gradazione del colore.

L'interazione fra un elettrone in un certo stato e tutti gli altri elettroni e nuclei positivi del materiale circostante è troppo complessa per poter essere visualizzata facilmente, ma con opportune approssimazioni e con l'ausilio di un calcolatore si può tracciare agevolmente l'andamento dell'energia di un elettrone in funzione della sua quantità di moto (*si veda l'illustrazione a destra*). La forma particolare della curva dipende dall'interazione. Inoltre, poiché la pendenza della curva energia-quantità di moto in un punto coincide con la velocità di un elettrone che si trovi nello stato rappresentato da quel punto, un grafico energia-quantità di moto permette di vedere chiaramente come la velocità dell'elettrone dipenda dalla struttura delle bande del materiale: una curva più ripida corrisponde a una velocità più elevata. (Ricordo che in meccanica quantistica la quantità di moto di un elettrone è molto diversa dalla quantità di moto classica, essendo funzione della lunghezza d'onda dell'elettrone. Anche un elettrone a riposo ha una lunghezza d'onda, perciò un elettrone con velocità nulla può avere una quantità di moto positiva.)

La velocità di un elettrone in moto entro un dato materiale, sotto l'influsso di un campo elettrico, può essere ricavata dalla pendenza della curva che esprime la relazione tra l'energia e la quantità di moto dell'elettrone. La forma della curva rispecchia le proprietà elettroniche del materiale. Quanto più forte è la sua pendenza, tanto più veloce è il moto dell'elettrone. Di conseguenza, due elettroni (*puntini*) in moto entro due materiali diversi possono avere la stessa quantità di moto, ma velocità diverse, come si vede dalle pendenze delle tangenti. Poiché in meccanica classica la quantità di moto viene definita come il prodotto della massa per la velocità, i fisici affermano che l'elettrone dotato di velocità più elevata ha una «massa efficace» minore.

Nonostante la loro natura quantistica, gli elettroni di un cristallo sono soggetti a due principi fondamentali della meccanica classica. Il primo è che la quantità di moto di un elettrone è modificata da una forza esterna (esercitata, per esempio, da un campo elettrico) agente su di esso; la quantità di moto di un elettrone soggetto a una forza costante, prodotta da un campo elettrico costante, aumenta a un tasso uniforme. Il secondo principio afferma che la variazione di energia di un elettrone è pari alla forza applicata moltiplicata per la distanza percorsa.

Se per effetto della stessa forza nella stessa unità di tempo due particelle raggiungono velocità diverse, in meccanica classica si dice che la particella più veloce ha massa minore. In effetti si adotta questa interpretazione nello studio del moto di un elettrone in un cristallo semiconduttore. Per una data forza applicata, si dice che quanto più elevata è la velocità di un elettrone, tanto minore è la sua «massa efficace». In termini delle curve energia-quantità di moto, dati due elettroni con la stessa quantità di moto, ma contenuti in materiali diversi, si può dire che la massa efficace dell'elettrone è minore nel materiale che manifesta la curva energia-quantità di moto più ripida.

Fino a quando la pendenza della curva energia-quantità di moto aumenta, cioè fino a quando la curva piega verso l'alto, l'elettrone accelera. Quando la curva comincia a piegare verso il basso (ossia quando l'elettrone si avvicina al suo massimo di energia), esso comincia a decelerare. Anzi, in corrispondenza del picco della curva la velocità dell'elettrone è nulla, perché in quel punto la curva ha pendenza nulla. Se la quantità di moto aumenta ancora per effetto di un campo elettrico, la velocità dell'elettrone diventa in realtà negativa, perché la curva piega verso il basso. In corrispondenza della porzione decrescente della curva si manifesta un paradosso: via via che la quantità di moto dell'elettrone aumenta in direzione del campo elettrico, aumenta anche la sua velocità nella direzione opposta! In breve, esso si muove in direzione opposta a quella del campo elettrico. Un modo per interpretare questo comportamento è di dire che esso ha acquistato una massa efficace negativa.

Questo comportamento piuttosto bizzarro dipende dal fatto che la gamma di energie contenute in una banda energetica è finita. Se continuasse a muoversi nella direzione del campo, l'elettrone continuerebbe ad acquistare energia dal campo e, quindi, ben presto la sua energia sarebbe maggiore di quella consentita dalla struttura delle bande. Quando l'elettrone comincia a muoversi nel verso opposto, in effetti, restituisce al campo l'energia che ne ha ricevuto.

Sotto il profilo matematico è conveniente combinare la lunghezza d'onda e la direzione di propagazione dell'elettrone in un'unica quantità, il vettore quantità di moto. In tal modo si può supporre che tutti i possibili vettori quantità di moto di tutti gli stati elettronici di un reticolo cristallino giacciano all'interno di un volume chiamato zona di Brillouin, le cui dimensioni non sono lunghezze, bensì quantità di moto secondo tre direzioni ortogonali. La forma della zona di Brillouin di un dato materiale dipende dalla sua struttura cristallina più che dalla sua composizione chimica. Per la maggior parte dei materiali semiconduttori, compresi il silicio e l'arseniuro di gallio, la zona di Brillouin ha la forma di un ottaedro i cui angoli siano stati mozzati, dando luogo a una figura con otto facce esagonali e sei facce quadrate.

Servendomi di un calcolatore, sono riuscito ad analizzare «fette» della zona di Brillouin passanti per il suo centro, cioè sezioni bidimensionali all'interno delle quali la quantità di moto può variare solo secondo due direzioni ortogonali. Circoscrivendo il problema in questo modo, ho potuto rappresentare ciascuna banda d'energia del materiale mediante una superficie, le cui variazioni d'altezza riflettono le variazioni di energia dell'elettrone. Una superficie energia-quantità di moto siffatta è analoga alle citate curve energie-quantità di moto, che rappresentano bande di energia in una sola dimensione dello spazio delle quantità di moto. Quindi la velocità e l'accelerazione di un elettrone sono date rispettivamente dalla pendenza e dalla curvatura locali della superficie.

Rappresentate in questo modo, le quattro bande di valenza del silicio e del-

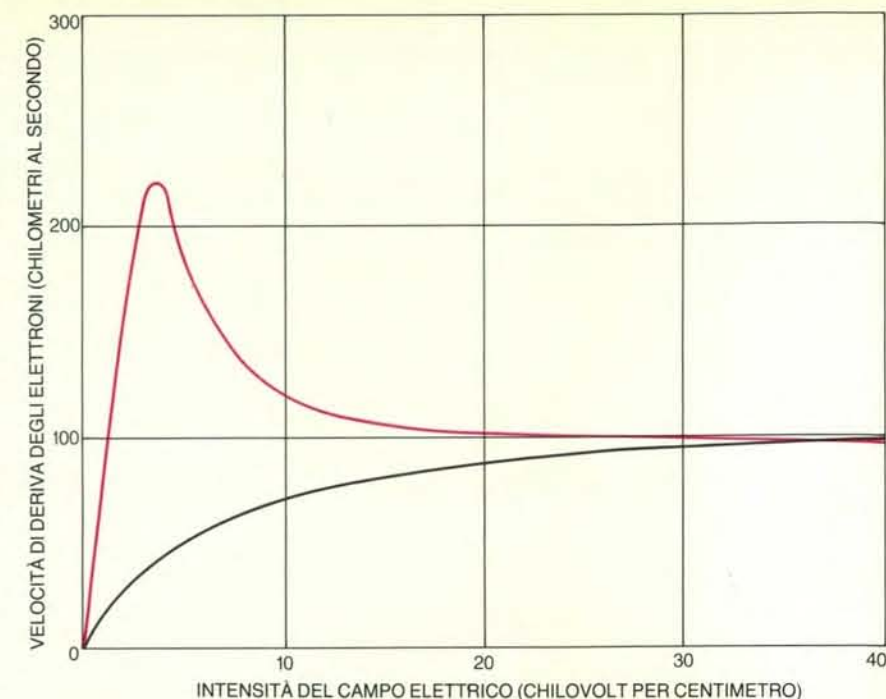
La dinamica di un elettrone nel silicio (*in alto*) e nell'arseniuro di gallio (*in basso*) può essere visualizzata tracciandone la quantità di moto entro una regione bidimensionale (*rappresentata dal tratteggio in nero*) contenuta nella zona di Brillouin del materiale; l'energia dell'elettrone viene riportata lungo una terza dimensione, ortogonale. Le superfici in rosso energia-quantità di moto si riferiscono agli elettroni di valenza, o di legame, degli atomi, che non possono condurre la corrente. Le superfici in blu corrispondono agli elettroni mobili della banda di conduzione, che in genere sono forniti dagli atomi delle impurezze contenute nel cristallo. La pendenza e la curvatura della superficie relativa alla banda di conduzione forniscono rispettivamente la velocità e l'accelerazione degli elettroni che trasportano la corrente. Questi elettroni tendono a occupare gli stati a energia minima sulla superficie. Nel silicio, gli «avvallamenti» in cui si raccolgono gli stati elettronici (su entrambi i lati del margine anteriore) hanno una curvatura ampia, e ciò indica che gli elettroni di bassa energia hanno una grande massa efficace. Nell'arseniuro di gallio, al contrario, il minimo della banda di conduzione cade nello stretto avvallamento centrale e ciò indica che gli elettroni di bassa energia di questo materiale hanno una massa efficace piccola e possono raggiungere velocità superiori agli elettroni del silicio quando vengono sottoposti al medesimo campo elettrico. Le illustrazioni sono state ottenute dall'autore applicando una tecnica ideata da Marvin L. Cohen dell'Università della California a Berkeley e da Thomas K. Bergstresser dei Sandia National Laboratories.

l'arseniuro di gallio sono simili, con la banda più bassa che presenta un minimo di energia nel centro della zona di Brillouin (corrispondente a una quantità di moto nulla), mentre le altre tre bande hanno un massimo di energia nel centro della zona. La somiglianza tra le bande di valenza dei due semiconduttori riflette la somiglianza tra i reticoli dei legami del silicio e dell'arseniuro di gallio.

Per il trasporto dei segnali elettrici, molti transistori comuni sfruttano assenze localizzate di elettroni nelle bande di valenza (invece che gli elettroni delle bande di conduzione). Queste «lacune» si ottengono rimuovendo elettroni dalla banda di valenza superiore, spesso mediante l'introduzione di atomi di impurezze aventi un elettrone in meno del corrispondente atomo del semiconduttore. Poiché le lacune elettroniche si comportano come quasi-particelle con carica positiva, esse possono fungere da trasportatori di carica nei transistori. Tuttavia, non sono adatte alle applicazioni che richiedono alte velocità: le bande di valenza superiori sono relativamente piatte sia nel silicio, sia nell'arseniuro di gallio, quindi le masse efficaci delle lacune sono piuttosto grandi e la loro velocità è intrinsecamente bassa.

Sopra le bande di valenza, in un semiconduttore, c'è un intervallo energetico proibito, o banda proibita, e ancora sopra vi sono le bande di conduzione. Poiché le bande di valenza sono di solito interamente occupate dagli elettroni di valenza, gli elettroni aggiunti al cristallo semiconduttore dagli atomi donatori impiantati devono occupare stati delle bande di conduzione, a partire dagli stati liberi a energia più bassa. Il trasporto ad alta velocità delle cariche ha luogo nella banda di conduzione più bassa.

Nel silicio lo stato a energia minima di questa banda si trova in corrispondenza di valori non nulli della quantità di moto: lontano dal centro della zona di Brillouin. In corrispondenza del minimo di energia i vettori quantità di moto possono assumere sei valori possibili (uguali in modulo, ma diversi per direzione). L'avvallamento intorno a ciascuno di questi stati è ampio e i suoi fianchi scendono gradatamente: ciò indica che gli elettroni accelerano lentamente e perciò hanno una massa efficace grande. Questa dipende dalla direzione del moto e varia da un valore pari circa alla massa di un elettrone libero, all'esterno del cristallo, a un quinto di questo valore.

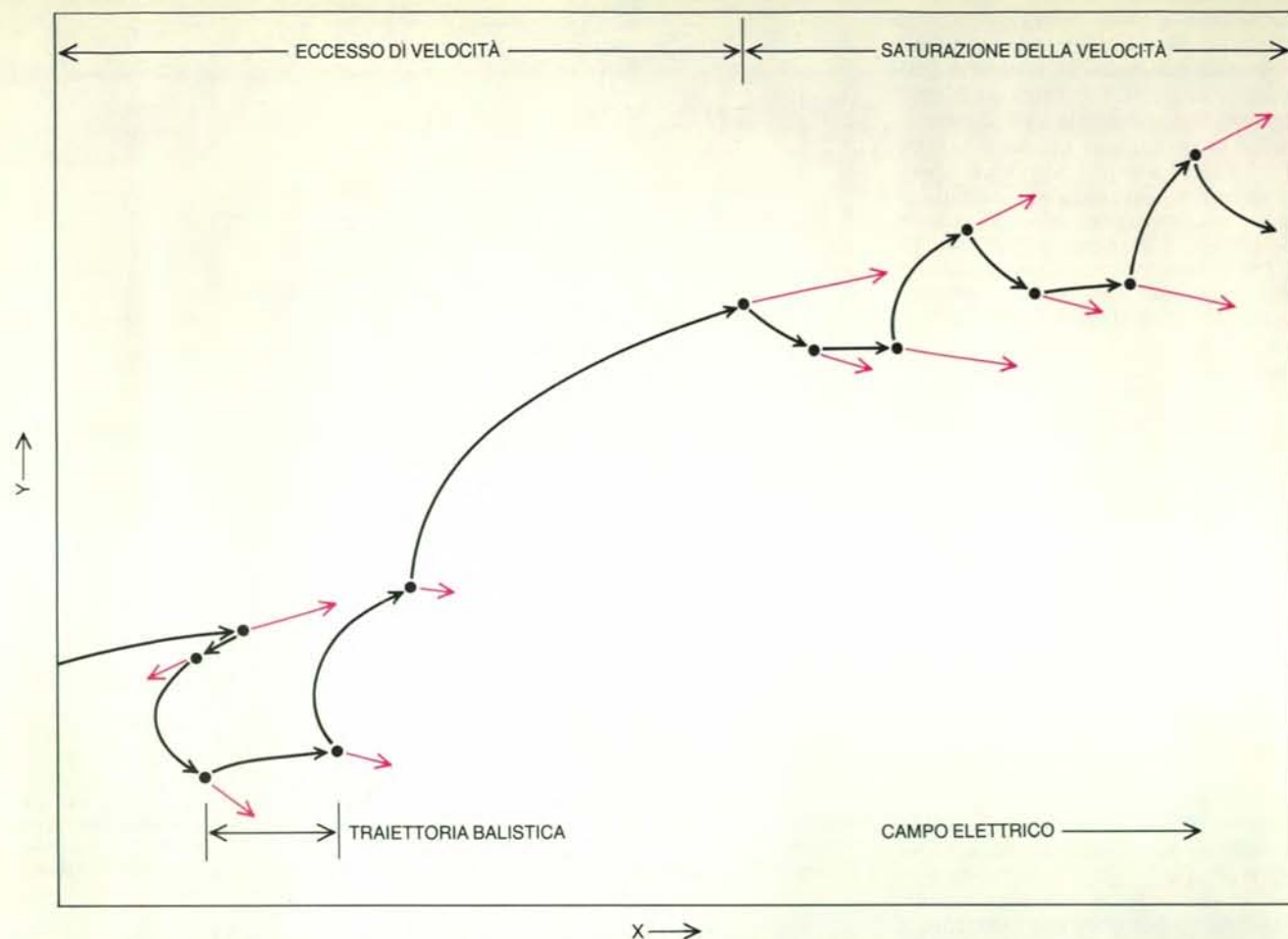
Nell'arseniuro di gallio la forma della banda di conduzione più bassa differisce da quella del silicio per le diverse interazioni che vi sono in questo composto fra elettroni e cariche nucleari. L'energia minima cade nel centro della zona di Brillouin e quindi vi è un solo minimo. Inoltre, l'avvallamento intorno a questo minimo è piuttosto stretto e quindi la massa efficace è piccola: solo 0,07 volte quella di un elettrone libero. L'esiguità della massa è uno dei motivi per cui gli

La velocità di deriva degli elettroni (la velocità media degli elettroni di conduzione) è riportata in funzione dell'intensità del campo elettrico per l'arseniuro di gallio (*in colore*) e per il silicio (*in nero*). Se l'intensità del campo è minore di 30 chilovolt per centimetro, la superiorità dell'arseniuro di gallio nei transistori ad alta velocità è evidentissima. Con campi più intensi, però, nell'arseniuro di gallio gli elettroni di conduzione diffondono fuori dallo stretto avvallamento centrale ed entrano in avvallamenti «satellite» più ampi alla superficie della banda di conduzione. Poiché gli avvallamenti satellite assomigliano agli avvallamenti della banda di conduzione del silicio, la velocità di deriva degli elettroni dell'arseniuro di gallio tende a quella degli elettroni del silicio.

elettroni si muovono più velocemente nell'arseniuro di gallio che nel silicio.

La massa efficace non è l'unico fattore che condiziona il moto degli elettroni in un cristallo semiconduttore, dato che il loro moto non è privo di ostacoli. Infatti gli elettroni urtano contro i difetti strutturali del reticolo cristallino, contro gli atomi di impurezze dotati di carica e contro i fononi, che sono vibrazioni termiche degli atomi del cristallo. Il loro moto quindi non è libero se non per periodi brevissimi (della durata di circa un decimo di bilionesimo di secondo) tra un urto e l'altro. In effetti, il moto degli elettroni in un semiconduttore ha la natura di un moto browniano, cioè aleatorio. Dato che l'applicazione di un campo elettrico fa accelerare gli elettroni fra un urto e l'altro in una direzione determinata, in media essi si spostano nella direzione di questo campo. Naturalmente quanto più un elettrone viene accelerato fra due urti consecutivi, tanto maggiore è la sua velocità media di deriva. Di conseguenza, nell'arseniuro di gallio la massa efficace più bassa degli elettroni fa aumentare la loro velocità di deriva.

Tuttavia la velocità di deriva aumenta anche all'aumentare degli intervalli fra gli urti. Anche sotto questo profilo la forma della banda di conduzione dell'arseniuro di gallio presenta dei vantaggi. La frequenza degli urti dipende dal numero di stati energetici liberi in cui può essere diffuso un elettrone dopo un urto. Nel silicio i sei ampi avvallamenti della banda di conduzione offrono molti stati del genere in prossimità del fondo della banda di conduzione. L'arseniuro di gallio, invece, possiede un numero relativamente piccolo di siti vicino allo stretto minimo della sua banda di conduzione.

Le conseguenze della diversa topografia della banda di conduzione nel silicio e nell'arseniuro di gallio si possono percepire con chiarezza tracciando la velocità di deriva di un elettrone in funzione dell'intensità del campo elettrico nei due semiconduttori. Se il campo elettrico è debole, la velocità di deriva nell'arseniuro di gallio cresce molto più rapidamente che nel silicio. (Si dice che nell'arseniuro di gallio la mobilità degli elettroni è maggiore che nel silicio.) La velocità degli elettroni nell'arseniuro di gallio raggiunge un valore massimo di circa 200 chilometri al secondo e poi, all'aumentare del campo elettrico, decresce fino a metà circa di questo valore. Nel silicio, la velocità si stabilizza intorno allo stesso valore per campi molto più forti.

La diminuzione della velocità di deriva degli elettroni osservata nell'arseniuro di gallio all'aumentare del campo elettrico (la cosiddetta mobilità differenziale negativa) può essere spiegata dalla topo-

La diffusione provocata dai fononi (*frecce in colore*) fa sì che in un materiale semiconduttore gli elettroni (*frecce in nero*), spostandosi nella direzione del campo elettrico applicato, seguano un cammino casuale. Fra un evento di diffusione e l'altro, l'elettrone segue una traiettoria balistica e quindi il moto di un elettrone che subisca pochi eventi di diffusione viene detto trasporto balistico. Nell'arseniuro di gallio, lo stato iniziale di un elettrone di conduzione è contenuto nell'avvallamento centrale della superficie energia-quantità di moto, dove la massa efficace è piccola. Un elettrone in questa situazione può essere accelerato da un campo elettrico tanto rapidamente da raggiungere un'elevata velocità di deriva. Ma questo «eccesso di velocità» non può essere mantenuto a lungo, perché, a velocità elevate, l'elettrone subisce facilmente la diffusione e passa dall'avvallamento centrale a uno degli avvallamenti satellite. Qui la massa efficace è più grande e quindi la velocità di deriva è molto minore: l'elettrone è molto più soggetto alla diffusione e si trova in una condizione detta di saturazione della velocità.

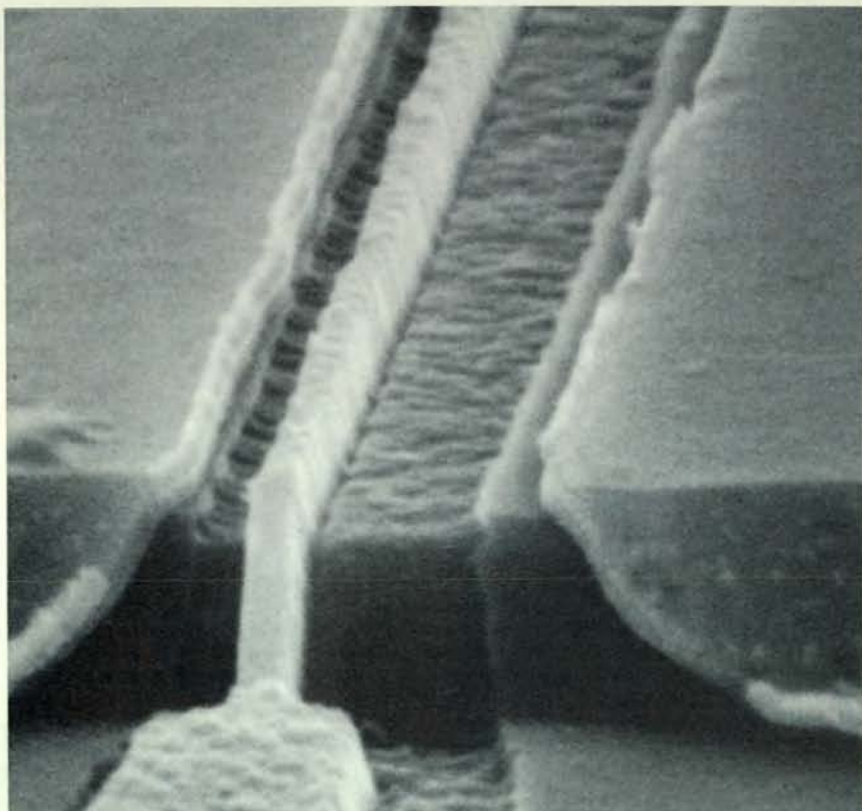
grafia della banda di conduzione esternamente allo stretto avvallamento centrale: essa assomiglia alla banda di conduzione del silicio in quanto vi sono parecchi avvallamenti in cui l'elettrone ha una massa efficace grande. Non deve, quindi, stupire che la velocità di deriva tenda, per campi elettrici più intensi, a quella osservata nel silicio. Via via che gli elettroni vengono accelerati da un campo abbastanza intenso, essi acquistano un'energia sufficiente a farli diffondere negli ampi avvallamenti «satellite» i cui stati a energia minima sono un po' più alti degli stati a energia minima dell'avvallamento centrale.

In certe condizioni, la diminuzione della velocità all'aumentare dell'intensità del campo elettrico provoca oscillazioni spontanee della corrente sia nell'arseniuro di gallio, sia in altri composti affini. Il fenomeno, su cui si basa il funzionamento di alcuni oscillatori a microonde allo stato solido, è l'effetto Gunn, dal nome di John B. Gunn, dell'International Business Machines Corporation, che scoprì queste oscillazioni nel 1963.

Finora ho discusso soltanto i fenomeni che rendono vantaggioso l'impiego dell'arseniuro di gallio nei FET tradizionali ad alta velocità. Tuttavia, a mano a mano che la tecnologia dei circuiti integrati all'arseniuro di gallio procede verso la produzione di massa, vengono sfruttati, per la progettazione di altri dispositivi allo stato solido, fenomeni più insoliti che si manifestano nell'arseniuro di gallio. Qui ne discuterò due: l'eccesso di velocità e il trasporto balistico.

Se la lunghezza della porta del FET è molto piccola, il campo elettrico dello strato attivo varia bruscamente in un breve tratto e un elettrone che si muova dalla sorgente al pozzo subisce un'improvvisa accelerazione. Anzi, questa accelerazione è così rapida che, prima di poter subire un gran numero di urti che assorbano la sua energia, l'elettrone può raggiungere velocità sostanzialmente superiori a quelle misurate su distanze più grandi in un campo elettrico uniforme. Il fenomeno, detto eccesso di velocità, fu scoperto da Jacques G. Ruch dei Bell Telephone Laboratories, grazie a simulazioni al calcolatore del trasporto elettrico. Queste simulazioni dimostrarono anche che l'effetto è molto più pronunciato nell'arseniuro di gallio che nel silicio a causa della minor massa efficace dell'elettrone e del tasso di diffusione inferiore, purché lo stato dell'elettrone si trovi nell'avvallamento centrale della banda di conduzione.

A causa di questo fenomeno, la velocità di deriva degli elettroni può addirittura raggiungere, nell'arseniuro di gallio, i 500 chilometri al secondo per tratti di pochi decimi di micrometro. Si pensa, anzi, che l'eccesso di velocità sia la causa principale delle buone prestazioni di velocità della maggior parte dei transistori di arseniuro di gallio, benché sia difficile darne una dimostrazione diretta.

Un fenomeno connesso, che si mani-

Questo transistore di arseniuro di gallio, costruito dalla Texas Instruments, Inc., ha un elettrodo porta (*cresta centrale*) largo soltanto 0,2 micrometri. L'elettrodo sorgente si trova ad appena 0,2 micrometri a sinistra della porta; il pozzo, a destra, è più lontano. Dispositivi di questo genere possono generare radiazioni elettromagnetiche di frequenza fino a 115 miliardi di hertz.

festa su distanze anche minori dell'eccesso di velocità, è il trasporto balistico, vale a dire il moto di elettroni senza perdite significative di energia. Anche questo effetto, che fu postulato da Michael S. Shur, dell'Università del Minnesota, e da Lester F. Eastman della Cornell University, è una conseguenza del fatto che gli urti non avvengono con continuità: un elettrone può percorrere una certa distanza senza subire collisioni. I dispositivi a semiconduttore possono essere resi così piccoli che la distanza che un elettrone deve percorrere è minore della distanza media fra gli urti.

Nell'arseniuro di gallio il trasporto balistico è stato osservato solo di recente nei cosiddetti transistori a elettroni caldi (si veda l'articolo *Elettroni balistici nei semiconduttori* di Mordehai Heiblum e Lester F. Eastman in «Le Scienze» n. 224, aprile 1987). Questi transistori hanno una lunghezza di transito piccolissima perché gli elettroni, invece di attraversare orizzontalmente lo strato attivo, come nel FET, attraversano verticalmente strati sottilissimi di semiconduttore. La lunghezza di transito dei dispositivi a trasporto balistico è ormai ridotta a 0,035 micrometri, ma l'effetto è utilizzato solo in dispositivi sperimentali. I transistori a trasporto balistico dovranno essere ancora perfezionati per poter competere con altri dispositivi ad alta velocità.

Nonostante il vantaggio costituito dalla maggior mobilità elettronica, i circuiti digitali all'arseniuro di gallio si affiancheranno a quelli di silicio, senza sostituirli. È improbabile che l'arseniuro di gallio venga impiegato nelle piastrine destinate ai calcolatori economici non specializzati, perché la produzione delle piastrine di silicio è molto meno costosa. Tuttavia molti sistemi digitali specializzati (per le telecomunicazioni e per alcune applicazioni militari) hanno componenti che devono funzionare alle massime velocità possibili. In questi sistemi, come in alcuni supercalcolatori che sono attualmente in fase di allestimento, saranno impiegati circuiti integrati all'arseniuro di gallio. Certi sistemi militari richiedono anche circuiti che funzionino bene a temperature elevate e in presenza di alti livelli di radiazioni. In queste condizioni i circuiti all'arseniuro di gallio si comportano meglio di quelli al silicio.

Nei circuiti analogici, tuttavia, la nuova tecnologia dell'arseniuro di gallio ha già avuto effetti importantissimi. Dato che i transistori di silicio sono troppo lenti per funzionare bene a frequenze superiori a 3 gigahertz (3 miliardi di cicli al secondo), i transistori di arseniuro di gallio sono diventati i dispositivi a semiconduttore di elezione per le applicazioni su gran parte della gamma delle microonde che si estende fino a 30 gigahertz. Se si potrà ridurre, con sufficiente affidabilità, la lunghezza della porta a un quarto di micrometro o meno, questi transistori si proporranno come la tecnologia dominante anche per i dispositivi funzionanti nella banda delle onde millimetriche, che va da 30 a 300 gigahertz e che finora è in gran parte inutilizzata.

I FET di arseniuro di gallio possono essere integrati con gli altri componenti necessari (come resistori, condensatori e induttori) su un'unica piastrina detta circuito integrato monolitico a microonde, o MMIC (*monolithic microwave integrated circuit*). Un MMIC, che ha un lato di appena qualche millimetro, può espletare funzioni che un tempo esigevano imgombranti circuiti costituiti da valvole termoioniche e da guide d'onda. Si potrebbero fabbricare molti MMIC su un'unica fetta di materiale, come oggi si fa per i circuiti logici di silicio, riducendo i costi di produzione. Grazie alle loro minuscole dimensioni e al loro piccolo peso, i MMIC consentiranno per esempio di costruire sistemi di radar a schiera in fase da montare su un aereo. Oggi, tali sistemi hanno le dimensioni di una casa.

I MMIC di arseniuro di gallio contribuiranno a ridurre i costi degli attuali sistemi a microonde, come i ricevitori per satelliti. I ricevitori domestici poco costosi costituiranno una parte importante dei sistemi di radiodiffusione diretta, che sono in via di allestimento in molti paesi. Questi sistemi di comunicazione su scala nazionale, in cui i segnali televisivi dovrebbero essere diffusi da satelliti anziché da antenne, funzioneranno a frequenze più elevate degli attuali sistemi di comunicazione via satellite. L'adozione di lunghezze d'onda più corte consentirà l'impiego di antenne a riflettore più piccole (di circa un metro di diametro), più economiche e facili da installare.

Le lunghezze d'onda corte alle quali potrebbero funzionare i trasmettitori e i ricevitori all'arseniuro di gallio consentirebbero anche nuove applicazioni. Per esempio, si potrebbero progettare proiettili «intelligenti» per carro armato, dotati di un radar in miniatura che li guidi sul bersaglio. Per poter entrare nel cannone del carro armato, l'antenna di un radar del genere dovrebbe essere piccolissima. Perché il radar abbia una risoluzione angolare abbastanza grande da guidare con precisione il proiettile, la lunghezza d'onda del segnale radar dovrà essere molto più piccola del diametro dell'antenna. Ciò significa che il radar deve elaborare onde millimetriche, che i FET di arseniuro di gallio possono generare e rilevare in modo efficiente.

I dispositivi all'arseniuro di gallio permetterranno certo di aprire a nuove applicazioni una vasta porzione inutilizzata dello spettro elettromagnetico. Per ora queste applicazioni sono soprattutto di carattere militare, ma via via che i costi per generare e rilevare i segnali a microonde diminuiranno, senza dubbio si scopriranno anche applicazioni civili.

La caccia alle gazzelle in Siria nell'Età della pietra

Lo scavo di Abu Hureyra, che documenta l'uccisione di interi branchi di animali a partire da 11 000 anni fa, rivela la persistenza nel Neolitico di strategie di caccia e raccolta anche dopo l'adozione dell'agricoltura

di Anthony J. Legge e Peter A. Rowley-Conwy

Uno dei problemi più importanti in campo archeologico è l'origine dell'economia agricola. Come e perché un modo di vita basato sulla caccia e sulla raccolta fu sostituito da un altro fondato sull'allevamento delle piante e degli animali? La questione è ovviamente complessa e nessun sito può singolarmente fornire la risposta. Tuttavia, vi sono siti archeologici che, grazie alla loro ubicazione e al periodo in cui furono occupati, possono aprire uno spiraglio sugli inizi della vita agricola.

Uno di questi siti è Tell Abu Hureyra, sulle sponde dell'Eufrate, nella Siria settentrionale. A partire da circa 11 000 anni fa, nel Mesolitico, la località fu occupata da un gruppo di cacciatori-raccoglitori e continuò a essere abitata, con una sola interruzione, fino al Neolitico avanzato. Poiché proprio in quest'epoca apparve l'agricoltura, Tell Abu Hureyra può fornire informazioni interessanti su tale processo di transizione.

Una delle scoperte più sorprendenti effettuate in questa località è che, forse ancora per un millennio dopo l'inizio della domesticazione delle piante, l'economia di caccia continuò a svolgere un ruolo fondamentale nella vita della comunità. Per tutto quel periodo la fonte principale di proteine animali rimase l'uccisione in massa delle gazzelle, effettuata durante le migrazioni dei branchi verso nord, all'inizio dell'estate. Sembra anzi che la strategia dell'uccisione in massa sia stata abbandonata solo in seguito al depauperamento dei branchi.

Tali inattese scoperte gettano nuova luce sugli inizi dell'agricoltura. Ad Abu Hureyra appare chiaro che alcuni fenomeni legati alla «rivoluzione agricola», come la coltivazione delle piante, possono imporsi molto tempo prima di altri (per esempio l'allevamento degli animali). In tale intervallo, non solo le tecniche di caccia e raccolta possono benissimo coesistere con quelle dell'agricoltura, ma la caccia può risultare più vantaggiosa della domesticazione per l'approvvigionamento di alimenti animali. Lezioni del genere vanno assimilate gradualmente, ma il lavoro condotto a Tell Abu Hureyra ha ormai cominciato a mostrare quanto sia stata complessa l'origine della vita agricola.

Il *tell* (o collinetta artificiale) di Abu Hureyra è un esteso rialzo che contiene vestigia di insediamenti umani accumulatesi in alcune migliaia di anni. La collinetta è in gran parte costituita di frammenti di muri di fango. Questo materiale forma un accumulo rilevante: il *tell* copre una superficie di 11 ettari e mezzo e gli strati archeologici hanno uno spessore di otto metri. Il deposito, quindi, può essere valutato intorno a un milione di metri cubi. Il *tell* fu scavato nel 1972 e nel 1973 da Andrew M. T. Moore, allora all'Università di Oxford (si veda l'articolo *Un villaggio agricolo preneolitico sull'Eufrate* di Andrew M. T. Moore in «Le Scienze» n. 134, ottobre 1979).

Uno scavo di emergenza

Si trattò di uno scavo di emergenza, in quanto il *tell* stava per scomparire: era infatti previsto che esso fosse sommerso nel 1974 dalla formazione del lago Assad, in seguito alla costruzione della diga di Tabqa, poco a valle sull'Eufrate. Dato che la collinetta appariva promettente sul piano archeologico, fu organizzato uno scavo rapido ma sistematico. Uno di noi (Legge) contribuì a pianificare i metodi di prelievo e fornì la strumentazione per il recupero di grandi quantità di resti organici, sia vegetali sia animali.

I resti organici raccolti nello scavo riflettono i modi di sussistenza degli abitanti del *tell*, che dipendono, a loro volta, assai strettamente dalla ubicazione di Abu Hureyra. La località si trova al margine della valle dell'Eufrate: da una parte c'è il fondovalle alluvionale, dall'altra una steppa arida, coperta di erba e arbusti. La zona riceve oggi non più di 200 millimetri di pioggia l'anno, giacendo così al limite dell'area in cui l'agricoltura può essere praticata senza irrigazione.

Il clima attuale non è molto diverso da quello che prevaleva verso il 9000 a.C., quando il sito fu fondato. (Tutte le date citate in questo articolo sono state ottenute con misure al radiocarbonio non «calibrate».) Durante il Mesolitico, il villaggio aveva forse 200-300 abitanti. Verso l'8000 a.C., sembra si sia verificato un periodo di abbandono, terminato verso il 7500 a.C. con la rioccupazione a opera di gruppi neolitici. Il villaggio nel Neolitico (lo stadio tradizionalmente associato all'inizio dell'agricoltura) era un insediamento più grande del precedente, con 2000 o 3000 abitanti.

I moderni metodi di scavo usati a Tell Abu Hureyra permettono di ripercorrere con notevole precisione i mutamenti nei modi di sussistenza avvenuti tra il Mesolitico e il Neolitico. Grandi campioni di sedimento furono trattati per flottazione e i reperti vegetali così ottenuti sono in corso di analisi da parte di Gordon Hillman dell'Istituto di archeologia dell'Università di Londra. Tutto il terreno scavato fu passato attraverso un setaccio con maglie di un centimetro, adatto a trattenere i reperti più grandi, fra cui le ossa. Su circa 60 000 frammenti ossei identificabili, abbiamo potuto analizzarne finora 40 000; il lavoro è finanziato dal Science and Engineering Research Council britannico.

Quali informazioni ci forniscono queste ossa? La prima conclusione è che la steppa era un ambiente molto più produttivo della valle alluvionale, come fonte di proteine animali. Per tutto il Mesolitico e per la prima parte del Neolitico, la specie di gran lunga più diffusa tra

Gli «aquiloni del deserto» sono antiche strutture per l'uccisione in massa di gazzelle, diffuse in Siria, Giordania, Arabia Saudita e nel deserto del Sinai. L'aquilone mostrato nella fotografia, che si trova a Dumayr, circa 30 chilometri a nord-est di Damasco, era probabilmente costruito in pietra. Le gazzelle venivano spinte nel corridoio fra i due muretti in basso fino al grande recinto, dove erano abbattute. Tre recinti più piccoli sono annessi a quello principale; uno di essi è stato sostituito da una struttura rettangolare (*in alto*). I piccoli cerchi lungo le pareti erano forse nascondigli per i cacciatori. L'immagine fu ottenuta nel 1930 da padre A. Poidebard, un pioniere della fotografia aerea in Siria.

quelle cacciate è la gazzella. Altre specie rappresentate sono l'onagro (che è molto simile all'asino selvatico), pecore, capre, suini, cervi e bovini selvatici, ma le ossa di gazzella costituiscono all'incirca l'80 per cento del totale. La specie più comune è la gazzella persiana, *Gazella subgutturosa*.

Normalmente le gazzelle non richiedono molto cibo fresco, come foglie e ramoscelli verdi, né molta acqua; poiché sono adattate alle regioni aride, vivono bene nella steppa. Suini e cervi, al contrario, hanno bisogno di bere spesso, e bovini, pecore e capre debbono bere almeno ogni pochi giorni. Tutte queste specie preferiscono quindi la valle fluviale alla steppa. La loro relativa scarsità a Abu Hureyra fa pensare che le specie di valle fossero cacciate fino a ridurne fortemente il numero durante le prime fasi dell'occupazione: in seguito fu la steppa a fornire la parte preponderante delle proteine animali.

La nostra analisi delle migliaia di ossa trovate nel *tell* non solo ha dimostrato che la gazzella era la specie predominante, ma ha fornito molte informazioni sul modo in cui queste piccole, veloci creature venivano uccise. Abbiamo notato che i resti di gazzella comprendono in abbondanza ossa e denti di animali molto giovani, ma non tutte le fasi di crescita sono rappresentate. Abbiamo quindi cominciato a esaminare i reperti nel tentativo di vedere quali fasi di crescita fossero presenti e quali no.

Un'eccellente fonte di notizie si sono rivelati i denti, il cui sviluppo segue uno schema ben noto. La gazzella persiana possiede alla nascita, o sviluppa ben presto, tre molari di latte su ciascuna metà della mandibola. I molari di latte durano un po' più di un anno, mentre dietro di essi si ha l'eruzione dei primi due molari permanenti. A poco più di un anno di età, i molari di latte sono sostituiti dai premolari permanenti. L'eruzione del terzo molare permanente avviene più o meno alla stessa epoca, completando l'apparato definitivo di sei denti su ciascuna metà della mandibola. Poiché le gazzelle si nutrono di vegetali coriacei, talvolta mescolati a granelli di sabbia, lo smalto della corona dei molari di latte subisce nel primo anno un'usura accentuata. L'usura dei denti di latte fornisce una buona indicazione dell'età di una giovane gazzella.

Per definire la distribuzione di età delle giovani gazzelle di Tell Abu Hureyra, ci siamo basati su misurazioni del terzo molare di latte. Esso è il più grande dei denti di latte e lo si riconosce facilmente perché possiede tre cuspidi, o superfici masticatorie. Misurando l'altezza della cuspide centrale, abbiamo trovato che i denti potevano essere divisi in due gruppi distinti: il primo gruppo non mostrava usura e aveva radici non del tutto formate, mentre nel secondo gruppo lo smalto della corona dentaria appariva fortemente abraso.

Purtroppo, *Gazella subgutturosa* allo stato selvatico è ormai quasi estinta nel Vicino Oriente, e quindi non c'è modo di trovare una spiegazione di tale andamento nell'area di studio. Ma la specie sopravvive nel Turkmenistan, in Unione Sovietica, dove, dal 1942 al 1947, V. G. Geptner del Museo di zoologia dell'Università statale di Mosca collezionò un gran numero di crani di gazzella. La collezione di Geptner è custodita in tale museo e uno di noi (Legge) ha esaminato i crani nel corso di una recente visita. Misurando l'altezza delle corone del terzo molare di latte, si è avuta una conferma di ciò che già sospettavamo: ad Abu Hureyra, i molari di latte privi di segni di usura erano quelli di animali neonati, mentre i denti fortemente abrasi provenivano da animali uccisi all'età di un anno.

Tell Abu Hureyra era un grande rialzo, o *tell*, sulle rive dell'Eufrate, nella Siria settentrionale. Le persone in primo piano sono beduini sedentarizzati, non più nomadi, che esercitano attività commerciali. Il *tell* era formato in gran parte dai resti dei muri di fango di un insediamento preistorico durato dal 9000 al 6000 a.C. Secondo gli autori, gli abitanti preistorici del luogo predavano i branchi di gazzelle per mezzo di un «aquilone del deserto». La costruzione in cima al dosso fa parte di un villaggio vicino. La cavità e i mucchi di terra visibili a sinistra testimoniano uno scavo archeologico condotto nei primi anni settanta. Successivamente il *tell* fu sommerso dal lago della diga di Tabqa, sull'Eufrate.

Tale risultato è stato confermato dall'esame dell'accrescimento e della saldatura delle ossa degli arti. Il calcagno, o osso del tallone, era adatto al nostro scopo. Come dice il nome colloquiale, il calcagno è parte del tallone, sebbene nelle gazzelle, che corrono sulle dita, esso sia alquanto distanziato dal terreno. Al pari del terzo molare di latte, si tratta di un osso facilmente identificabile, che presenta caratteri di accrescimento ben riconoscibili. Uno di questi caratteri è il fatto che, durante il primo anno di vita, il corpo dell'osso si allunga e si ispessisce considerevolmente; inoltre, all'età di circa 14 mesi, un'apofisi (una piccola protuberanza in cui si inserisce un tendine, che alla nascita si presenta come elemento osseo separato) si salda all'osso principale.

Le migrazioni delle gazzelle in Giordania e in Siria seguivano una direttrice nord-sud. In base alla distribuzione degli «aquiloni del deserto» e ad altri indizi, gli autori hanno ricostruito due vie (*in colore*). Nella tarda primavera, le gazzelle si spostavano a nord, dove nascevano i piccoli; alla fine dell'estate i branchi tornavano a sud. La via occidentale è ipotetica;. quella orientale termina presso Tell Abu Hureyra: qui i branchi erano cacciati al loro arrivo nei mesi di aprile e maggio.

La distribuzione di età

Le ossa delle giovani gazzelle mostravano una buona corrispondenza con la distribuzione osservata nei molari di latte. Un gruppo di calcagni comprendeva esemplari di piccole dimensioni con l'apofisi separata: ovviamente doveva trattarsi di neonati. Un secondo gruppo, sebbene l'apofisi fosse ancora separata, comprendeva ossa di dimensioni quasi uguali a quelle degli adulti: si trattava di animali di un anno. Il gruppo più numeroso era poi costituito da ossa grandi con apofisi saldata: questi calcagni provenivano da gazzelle adulte. In seguito, lo studio dei denti permanenti ha confermato la presenza di adulti di tutte le età accanto ai neonati e ai soggetti di un anno.

Come si spiega che vi sia una divisione delle ossa e dei denti in classi di età di «neonati», «soggetti di un anno», e «adulti»? La risposta dipende in parte dal fatto che il ciclo biologico di un branco di gazzelle è altamente sincronizzato, ossia tutte le nascite hanno luogo in un breve periodo. Se il branco è cacciato anno dopo anno nello stesso momento stagionale, gli animali uccisi mostreranno una distribuzione di età caratteristica, comprendente gruppi di giovani animali a intervalli di un anno. È verosimilmente quanto deve essere accaduto a Tell Abu Hureyra: la caccia alle gazzelle costituiva un'occupazione stagionale. La presenza, fra le prede, di animali neonati indica che il momento della caccia cadeva poco dopo l'epoca in cui le femmine partorivano. Notizie tramandate dai primi viaggiatori europei indicano che le gazzelle della zona di Abu Hureyra figliavano tra la fine di aprile e l'inizio di maggio.

Non solo ci trovavamo di fronte ad abbattimenti stagionali, ma appariva evidente che obiettivo della caccia era il branco, non le singole gazzelle. La distribuzione delle ossa trovate nel *tell* non può essere derivata dalla caccia a singoli animali. Quando dei cacciatori tendono l'agguato a un branco e scelgono singole gazzelle, essi puntano ad animali giovani in perfette condizioni, spesso soprattutto maschi. Ma ad Abu Hureyra abbiamo trovato ossa di tutte le classi di età, dagli individui più giovani ai più vecchi: una distribuzione risultante, senza alcun dubbio, da una tecnica di caccia che permetteva di catturare l'intero branco in un'unica soluzione.

Metodi di uccisione in massa sono ben noti, sia nella preistoria, sia presso i cacciatori-raccoglitori di epoca storica, come gli indiani delle praterie, nell'America Settentrionale. Gli indiani utilizzavano vari metodi per catturare branchi di bisonti e di antilocapre (*Antilocapra americana*), ma i diversi sistemi avevano un denominatore comune abbastanza semplice: il branco era spinto verso una recinzione o una grande buca, opportunamente mascherata, all'interno della quale si poteva procedere all'abbattimento nella maniera più efficiente. In certi casi, si spingevano gli animali attraverso un passaggio obbligato, che conduceva a una destinazione scelta dai cacciatori.

Se sia vantaggioso o no per un gruppo di cacciatori-raccoglitori adottare questi metodi dipende da molte considerazioni. L'abbattimento in massa è da considerarsi una strategia valida soprattutto in aree aperte nelle quali i grandi branchi offrono una facile preda. Durante la battuta di caccia occorrono numerosi uomini per dirigere i branchi e far funzionare le trappole nei luoghi prescelti. Questi tre criteri - terreno aperto, grandi branchi e un sufficiente numero di uomini - furono soddisfatti a Tell Abu Hureyra nel Mesolitico e all'inizio del Neolitico.

Gli «aquiloni del deserto»

Benché le strategie di uccisione in massa siano documentate così ampiamente nell'America Settentrionale, nel Vicino Oriente esse non hanno ricevuto molta attenzione fino a tempi recenti. Negli ultimi anni, però, gli archeologi hanno incominciato a rendersi conto che strutture per l'uccisione di branchi di gazzelle sono molto comuni in Giordania, Siria, Arabia Saudita e nel deserto

del Sinai. Tali strutture rimasero in uso per lungo tempo, tanto che viaggiatori europei ne videro ancora in funzione nel XX secolo, ma sulle prime né la loro diffusione né il loro scopo apparivano chiari. L'ampia distribuzione delle strutture fu riconosciuta solo quando nacquero le rotte postali aeree e i piloti si diedero a fotografare i siti archeologici dall'alto. Nel 1929, il comandante di gruppo L. W. B. Rees, un pilota della linea Cairo-Baghdad, pubblicò un primo resoconto sulla rivista «Antiquity». Rees diede alle strutture il nome di «aquiloni del deserto», per l'apparente somiglianza, dall'alto, con gli aquiloni.

Gli aquiloni del deserto variano per forma e grandezza, ma hanno alcuni elementi in comune. Di solito, vi è un recinto centrale che può avere un diametro variabile da pochi metri fino a 150 metri. Il recinto ha un ingresso ristretto, dal quale divergono due muretti di pietre, che possono prolungarsi nel deserto per diversi chilometri. In alcune strutture, il recinto e i muretti sono realizzati con grandi lastre di pietra poste a coltello, in altri casi consistono semplicemente di blocchi ammucchiati.

Rees e gli osservatori che lo seguirono negli anni trenta e quaranta erano dei militari. Essi diedero una interpretazione militare degli aquiloni, scambiando i recinti per strutture destinate alla difesa di uomini e bestiame. Da allora è però diventato chiaro che gli aquiloni servivano in realtà per abbattimenti in massa. L'estremità aperta della V formata dai muretti è situata molto spesso in prossimità di piccole valli, attraverso le quali si spostavano i branchi di gazzelle. I cacciatori deviavano il branco verso l'apertura e spingevano gli animali terrorizzati dentro il recinto murato destinato all'uccisione, nascosto non di rado dietro il ciglio di una bassa collina.

La gazzella persiana, *Gazella subgutturosa*, era la preda principale dei cacciatori di Abu Hureyra. Il maschio (*in basso*) ha un'altezza di 60 centimetri al garrese. In Giordania e Siria la gazzella persiana è estinta, allo stato selvatico, per la caccia indiscriminata condotta in età recente.

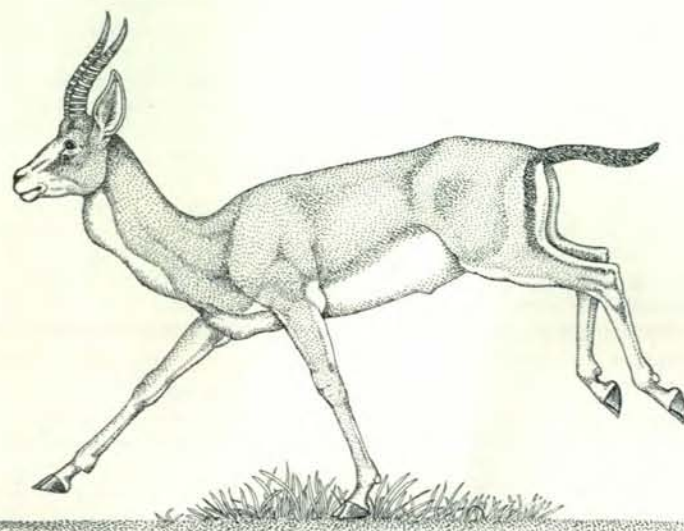
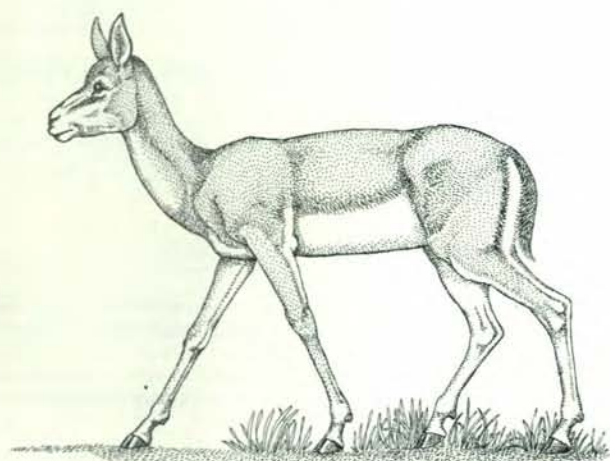
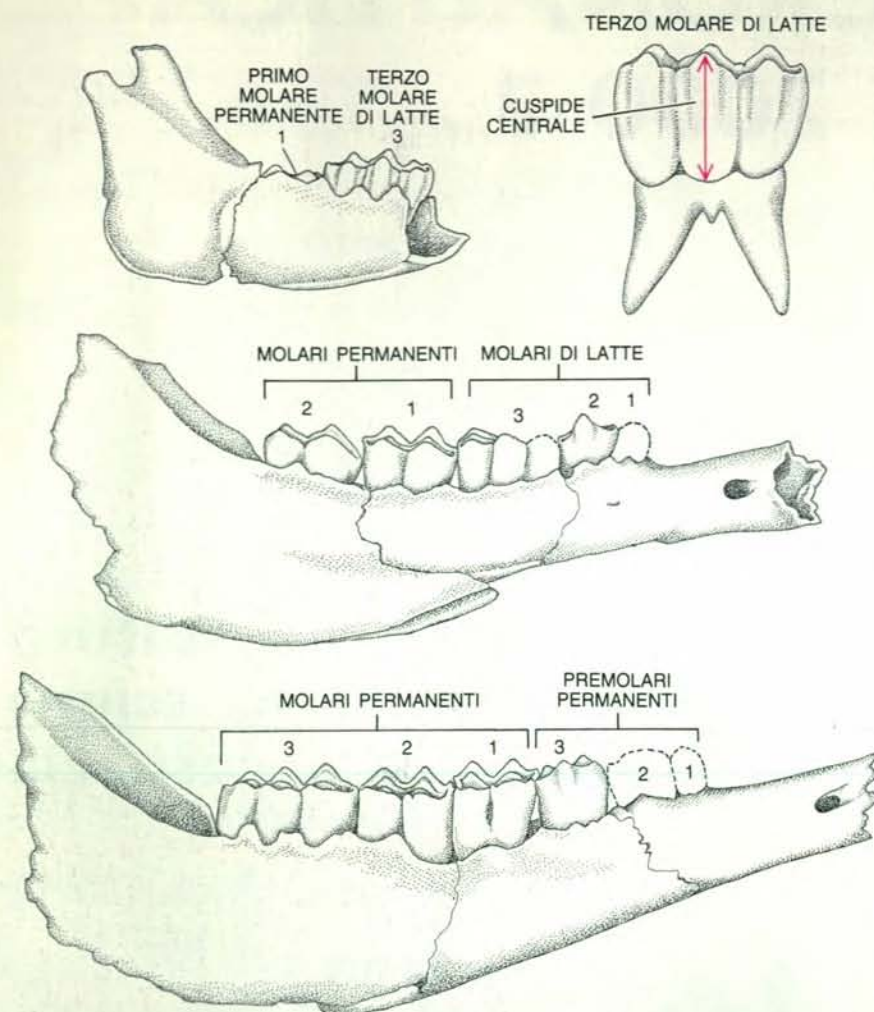
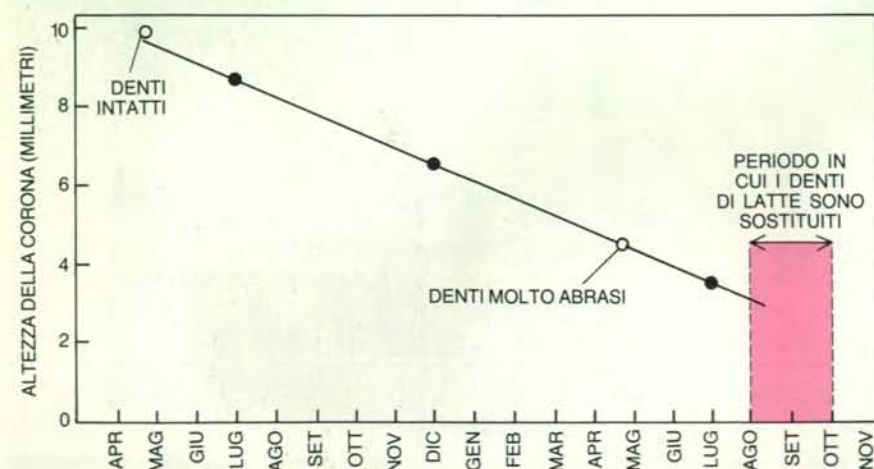
Secondo recenti lavori archeologici e notizie di viaggiatori, un aquilone poteva funzionare in due modi, tenuto conto della sua ubicazione e delle caratteristiche di costruzione. In Giordania, il materiale utilizzato prevalentemente è dato da blocchi di basalto derivati dalle lave vulcaniche che ricoprono estesamente la parte settentrionale e orientale del paese. In questo caso, il recinto per l'abbattimento presenta, nello spessore dei muretti di pietre, piccoli nascondigli rotondi simili alla parte visibile di un pozzo per acqua. Svend W. Helms dell'Istituto di archeologia dell'Università di Londra ha ipotizzato che in tali nicchie si nascondessero i cacciatori pronti a colpire le gazzelle intrappolate con arco e frecce.

In Siria, dove non vi sono estesi affioramenti di antiche lave, si preferiva un'altra strategia, scavando trappole a pozzo all'esterno del recinto principale. J. L. Burckhardt, che percorse il Vicino Oriente nel XIX secolo, vide aquiloni con trappole a pozzo in funzione presso il villaggio siriano di Al-Qaryatayn. Egli riferisce che il muretto del recinto aveva dei varchi, saltando i quali le gazzelle impaurite finivano dentro le trappole. Le gazzelle ferite nella caduta erano rapidamente finite e macellate: si rimuovevano subito le ossa degli arti e la carne era salata e seccata per la conservazione. Le scorte sarebbero state senz'altro consumate nel giro di un anno.

Sembra evidente che una struttura del genere debba essere stata impiegata a Tell Abu Hureyra durante i millenni che videro il grande accumulo di ossa. Non solo, ma l'aquilone doveva essere situato abbastanza vicino all'abitato, a giudicare dalla collezione archeologica, che comprende un buon assortimento di tutte le ossa dello scheletro. Le ossa degli arti non sono scarsamente rappresentate, come avverrebbe se l'aquilone fosse stato ubicato lontano dal villaggio. Le notizie storiche indicano infatti che, quando le gazzelle erano parzialmente macellate sul posto prima del trasporto, le ossa degli arti venivano abbandonate sul luogo dell'abbattimento.

Le migrazioni delle gazzelle

Non abbiamo però ancora prove dirette di una struttura per l'uccisione in massa. Allorché il *tell* fu scavato, nessun archeologo si aspettava di scoprire che la caccia fosse praticata su scala così grande in una comunità già caratterizzata dai primordi della vita agricola. Perciò la zona non fu ispezionata alla ricerca delle tracce di un aquilone. Ma non tutto è perduto: se il *tell* è sommerso, gran parte del territorio adiacente non lo è. Tenendo conto della morfologia del terreno, si può ipotizzare con una certa probabilità quale fosse la posizione dell'aquilone. In tempi brevi, intendiamo tornare ad Abu Hureyra per cercare di trovarne i resti.

Anche in mancanza di resti materiali dell'aquilone, è comunque possibile de-

I denti delle gazzelle danno indicazioni sull'età degli animali uccisi a Tell Abu Hureyra. Le mandibole nell'illustrazione provengono dal sito. Alla nascita, l'animale ha tre molari decidui o «di latte». La mandibola in alto, appartenente a un neonato, conserva il terzo molare di latte, dietro il quale si vede il primo molare permanente. Nel primo anno di vita si ha l'eruzione dei molari permanenti. La mandibola al centro appartiene a un soggetto di un anno e ha tre molari di latte e due molari permanenti. Verso l'età di 14 mesi, i molari di latte sono sostituiti dai premolari permanenti: l'esemplare in basso, di un adulto, presenta i tre premolari e i tre molari permanenti.

L'usura del terzo molare di latte indica la stagione della caccia alle gazzelle. Nel primo anno di vita, la corona del terzo molare di latte subisce una rapida abrasione. Il tasso di usura è stato calcolato da misurazioni su crani di gazzelle attuali (*cerchietti pieni*). Gli esemplari di terzo molare di latte di Tell Abu Hureyra (*cerchietti vuoti*) vengono da neonati o da soggetti di un anno. Quindi le gazzelle dovevano essere cacciate in aprile e maggio, quando i piccoli erano appena nati.

durre molti dettagli su come l'uccisione in massa si inserisse nel quadro ecologico e sociale della comunità. Un aspetto di tale quadro è la migrazione stagionale dei branchi di gazzelle. Come mostrano i dati archeologici, le gazzelle erano uccise a Tell Abu Hureyra in un breve periodo all'inizio dell'estate. In tale periodo, probabilmente i branchi raggiungevano il limite settentrionale della loro migrazione annuale.

Come si è detto, la gazzella persiana è eccezionalmente bene adattata alle condizioni aride. Di solito, l'animale ingerisce i liquidi di cui ha bisogno mangiando vegetali secchi ricoperti di rugiada. Ma c'è almeno un'eccezione alla regola: le femmine che debbono produrre latte hanno bisogno di acqua o di foglie, steli e ramoscelli verdi. Per soddisfare questo bisogno, i branchi di gazzelle si spostavano a nord in primavera, lasciando il deserto. Quando raggiungevano l'ambiente relativamente più umido presso la valle dell'Eufrate, le femmine gravide partorivano i piccoli. I branchi passavano alcuni mesi nel loro territorio settentrionale prima di puntare di nuovo a sud, in luglio.

La topografia della regione, i resoconti dei viaggiatori e la distribuzione degli aquiloni del deserto fanno pensare che i branchi di gazzelle seguissero, nel loro spostamento annuale da nord a sud, due vie principali. Un itinerario occidentale - abbastanza ipotetico - seguiva la valle che si estende dalla zona di Damasco verso nord-est fino al distretto di Homs, a ovest di una bassa catena montuosa chiamata Gebel Abu Rujmayn. Gli aquiloni visti da Burckhardt sono situati lungo questa via.

L'altro itinerario, assai meglio documentato, passa a est del Gebel Abu Rujmayn. Tra questa catena e quella vicina chiamata Gebel Bishri si apre una bassa insellatura già sfruttata dalle carovane sul percorso da Aleppo al golfo Persico. I resoconti storici dimostrano che anche i branchi di gazzelle usavano questo passaggio. Appena a nord della sella sta il villaggio di At-Taybe, dove, nel 1604, il viaggiatore portoghese Pedro Teixeira vide in uso aquiloni per la caccia alle gazzelle. Questi aquiloni erano diversi da quelli osservati in altre zone della Siria oppure in Giordania, perché, invece di muretti, per dirigere il branco essi avevano file di bandierine di cenci, sventolanti su paletti. Le gazzelle, spaventate dall'agitarsi delle bandierine, erano spinte verso il recinto tra i filari di paletti.

Nonostante la diversa forma dei «muretti», gli aquiloni visti da Teixeira venivano fatti funzionare più o meno come il tipo a muri di pietre e i due tipi, in effetti, coesistevano nella stessa regione. Una ricognizione svolta negli anni trenta da padre A. Poidebard, un pioniere della fotografia aerea in Siria, rivelò l'esistenza di aquiloni con muretti di pietre a sud di At-Taybe.

Ci siamo interessati in particolare a questa concentrazione di aquiloni perché si tratta del gruppo più vicino ad Abu Hureyra. La sua esistenza conferma che l'uccisione in massa delle gazzelle era una strategia di sussistenza praticata nella regione di Abu Hureyra, almeno in epoca storica. Sebbene i branchi di gazzelle fossero cacciati in più punti lungo la loro via di migrazione, il punto più settentrionale in cui la caccia sistematica era possibile era proprio Tell Abu Hureyra. Raggiunto questo sito, i branchi si separavano per l'estate in gruppi più piccoli e la strategia dell'uccisione in massa non poteva più essere impiegata fino al momento in cui i branchi si ricomponevano per migrare a sud nel mese di luglio.

Tra caccia e domesticazione

Il quadro finora ricostruito offre importanti informazioni sulle strategie economiche preagricole a Tell Abu Hureyra e anche sull'origine dell'agricoltura nella zona. Se si considerasse soltanto la pratica dell'uccisione in massa, Tell Abu Hureyra dovrebbe essere visto come un accampamento stagionale legato all'attività di caccia alle gazzelle. Il resto della documentazione, tuttavia, permette di affermare che il sito fu occupato in permanenza, fin dall'inizio.

Un aspetto non trascurabile di tale documentazione è l'imponenza del *tell* stesso, le cui dimensioni implicano quasi certamente una vita sedentaria di villaggio. Inoltre, gli avanzi vegetali dai livelli mesolitici del *tell* indicano che veniva raccolta una grande varietà di potenziali alimenti vegetali e che le stagioni di raccolta di tali piante abbracciavano gran parte dell'anno o l'anno intero.

Ad Abu Hureyra, nel Mesolitico, erano già presenti aspetti dell'economia agricola? Non è impossibile che lo fossero. Fra i resti vegetali che Hillman sta studiando, si trovano chicchi carbonizzati di frumento. Tuttavia, si tratta di una forma selvatica del tipo di frumento chiamato monococco o *einkorn*, della cui coltivazione non abbiamo prove dirette. Indicazioni di maggiore rilievo vengono dal ritrovamento di ossa di capra e di pecora. Le datazioni al radiocarbonio mediante acceleratore di particelle, effettuate all'Università di Oxford, hanno confermato che alcune delle ossa di questi animali possono essere riferite al Mesolitico. Ma si trattava di animali addomesticati, allevati allo scopo di disporre di pelli per indumenti, di latte e di carne, oppure di animali selvatici, cacciati come fonte di cibo al pari delle gazzelle?

La questione è complessa. Oggi le pecore o le capre selvatiche non frequentano la steppa di bassa quota intorno ad Abu Hureyra. Questi animali rifuggono da zone del genere, preferendo altitudini più elevate e - specialmente le capre - terreno più ripido. Ma questi ambienti di altitudine possono essere semplicemente i soli luoghi in cui capre e pecore sono sopravvissute alla pressione ecologica dell'uomo. Non è da escludere che in epoche passate capre e pecore selvatiche possano essere vissute vicino ad Abu Hureyra. In ogni caso, non è facile distinguere le ossa di capre e pecore selvatiche da quelle dei loro discendenti domestici, sulla sola base delle dimensioni e della forma.

Stabilità e mutamento

Se si ipotizza che capre e pecore selvatiche fossero presenti nella steppa di Abu Hureyra, si deve tuttavia notare anche che queste specie sono rare in altri siti mesolitici del Vicino Oriente. Non c'è nessun altro *tell* scavato con metodi moderni che abbia fornito strati contemporanei a quelli più profondi di Abu Hureyra. I siti della regione risalenti a un'epoca più antica, tra 50 000 e 10 000 anni a.C., sono piccoli e non presentano la struttura a *tell*, che caratterizza i siti posteriori, come Tell Abu Hureyra. In questi siti più antichi pecore e capre sono molto rare: fino a questo momento tutti i siti hanno fornito, in totale, non più di qualche centinaio di ossa di pecora. Sembra dunque che ad Abu Hureyra stesse avvenendo qualcosa di insolito, probabilmente connesso a tecniche di domesticazione primordiali.

Ma anche se pecore e capre erano state addomesticate, il loro significato nell'economia animale complessiva della località, in epoca mesolitica, era rimasto relativamente modesto. Nell'insieme, pecore e capre rappresentano soltanto il 10 per cento delle ossa trovate nei livelli mesolitici, rispetto - come abbiamo visto - all'80 per cento delle ossa di gazzella. È notevole che tale andamento continui nella prima parte del Neolitico. L'uccisione stagionale delle gazzelle veniva ancora praticata e le loro ossa costituivano sempre il 70 o l'80 per cento del totale. Eppure in quest'epoca la coltivazione delle piante si era già diffusa nel Vicino Oriente e - come ha dimostrato il lavoro di Hillman - l'agricoltura era praticata su vasta scala a Tell Abu Hureyra.

Oltre alla coltivazione delle piante, sembra fosse già attestato il secondo grande aspetto di un'economia agricola, la domesticazione degli animali. All'inizio del Neolitico le ossa riferibili a capre e pecore sono sufficientemente numerose da permettere la misurazione dell'altezza delle corone, nel terzo molare di latte, come è stato fatto per le gazzelle. Le misure indicano che, a differenza delle gazzelle, capre e pecore non erano uccise con ritmo stagionale, ma venivano macellate durante tutto l'anno. È ciò che ci si aspetta in un insediamento agricolo in cui capre e pecore sono domesticate in maniera sistematica e permanente. Tuttavia, l'importanza di queste specie rimane bassa e le gazzelle continuano a essere predominanti.

L'accrescimento del calcagno delle gazzelle conferma il ritmo stagionale della caccia. Nel primo anno di vita, il calcagno (*nel particolare*) si ispessisce notevolmente; a circa 14 mesi, si ha la saldatura dell'apofisi (la protuberanza ossea nella quale si inserisce il tendine). I calcagni provenienti da Tell Abu Hureyra sono di tre tipi. Quelli di neonati (*in grigio*) sono sottili e hanno l'apofisi separata; quelli di soggetti di un anno (*in bianco*) sono più spessi, ma l'apofisi non è ancora saldata. I calcagni di adulti (*in colore*) sono spessi e la saldatura dell'apofisi è completa. La ripartibilità in gruppi dei reperti implica che l'abbattimento degli animali era stagionale.

Questo fatto è interessante o, per meglio dire, sorprendente. Come è possibile che capre e pecore, così strettamente associate agli esordi della domesticazione, mantengano un ruolo minore molto tempo dopo l'affermarsi della vita agricola? Una spiegazione va cercata nella presenza di cospicui branchi di gazzelle, che fornivano annualmente una sicura base di risorse. Quando la provvista di carne per tutto l'anno può essere fatta in poche settimane, ogni primavera, non vi sono molti incentivi a intraprendere la difficile pratica dell'allevamento su vasta scala.

Questa situazione favorevole non durò per sempre. A un certo punto, nel settimo millennio a.C., l'economia animale a Tell Abu Hureyra subì un cambiamento brusco e profondo: capre e pecore sostituirono le gazzelle quanto a importanza relativa. In breve tempo, le gazzelle declinarono fino a fornire appena il 20 per cento delle ossa, mentre capre e pecore crescevano di importanza fino a raggiungere quel livello dell'80 per cento che era già stato occupato dalle gazzelle.

Sembra che questa radicale trasformazione sia da ricollegare ad avvenimenti verificatisi molto più a sud. È probabile che, all'epoca in cui la cattura delle gazzelle cominciò ad essere praticata ad Abu Hureyra, gli aquiloni del deserto non fossero ancora comuni in Siria e in Giordania. In effetti, sono stati rinvenuti in queste zone ben pochi siti archeologici del 9000 a.C., contemporanei all'insediamento mesolitico di Tell Abu Hureyra. Verso la metà del periodo neolitico, però, gli aquiloni erano diventati assai più numerosi.

Le prime ricognizioni aeree, integrate dal recente lavoro sul campo di Alison Betts e del suo collega Helms, hanno portato alla scoperta di una grande concentrazione di aquiloni nella Giordania settentrionale. Oltre a essere costruiti piuttosto vicini tra loro, questi aquiloni sono anche collegati l'uno all'altro. I muretti destinati a convogliare i branchi verso l'interno di un recinto sono connessi a quelli di un recinto contiguo, formando grandi barriere che si estendono, talvolta, per decine e decine di chilometri attraverso la pianura. Con questa disposizione, gli aquiloni devono avere avuto effetti micidiali sui branchi di gazzelle.

Le percentuali relative delle varie specie animali presenti a Tell Abu Hureyra, indicate dalla larghezza dei settori in colore, rivelano un mutamento radicale nel settimo millennio a.C. Dal 9000 a.C. fino a poco dopo il 6500 a.C., le ossa di gazzelle rappresentano l'80 per cento e quelle di pecore e capre il 10 per cento del totale. A partire da quest'epoca, la situazione si capovolge: pecore e capre salgono a oltre il 60 per cento, mentre le gazzelle diminuiscono fino al 20 per cento. Sembra che il mutamento sia stato causato dalla decimazione dei branchi di gazzelle.

Siti del genere sono databili con difficoltà, in quanto sono ubicati su superfici rocciose e i sedimenti - strumento di datazione di primaria importanza - vi si accumulano lentamente. Pare comunque dimostrato che alcuni di questi aquiloni siano stati costruiti più o meno all'epoca in cui i residenti di Abu Hureyra cominciarono a trarre sostentamento più dalle capre e dalle pecore che dalle gazzelle. La Betts ha descritto utensili di selce riferibili approssimativamente a tale data, provenienti da aquiloni della Giordania orientale. Abbiamo esaminato materiali a essi correlabili trovati in aquiloni della zona di Azrak, un'oasi che si trova 60 chilometri a est di Amman, in siti ispezionati da A. N. Garrard del British Institute of Archaeology and History ad Amman.

Utensili di pietra neolitici sono stati infine segnalati in alcuni aquiloni individuati nell'Arabia Saudita settentrionale. Essi, in genere, provengono dall'interno dei recinti, ma talvolta sono stati rinvenuti anche nelle nicchie-nascondiglio nelle quali i cacciatori dovevano probabilmente appostarsi.

Questo complesso di documenti suggerisce che, nel settimo millennio a. C., gli aquiloni del deserto si siano diffusi rapidamente nelle steppe della Giordania settentrionale. Naturalmente, la cattura di un branco mediante una strategia di uccisione in massa può causare l'abbattimento di un numero eccessivo di animali. A quando pare, questo è proprio ciò che accadde. Nel Mesolitico e all'inizio del Neolitico, gli abitanti di Tell Abu Hureyra erano fra i pochi gruppi che cacciavano gazzelle con metodi di uccisione in massa. Nel periodo successivo, la diffusione degli aquiloni nella Giordania settentrionale può avere cominciato a esercitare una pressione eccessiva sui branchi. La conseguenza può essere stata la diminuzione della popolazione di gazzelle e forse, addirittura, lo sconvolgimento totale del regime migratorio di questi animali. La diminuita disponibilità di gazzelle, a sua volta, potrebbe aver costretto gli abitanti di Abu Hureyra a ripiegare sull'allevamento di capre e pecore.

Una «rivoluzione neolitica»?

Così, con un ritardo di un millenio, l'allevamento di animali su larga scala venne a integrare la coltura dei vegetali, e il villaggio di Tell Abu Hureyra entrò pienamente nella fase agricola. Uno de-

Le oscillazioni caccia-allevamento: contributi della ricerca zooarcheologica italiana

L'articolo di Legge e Rowley-Conwy si inserisce in una corrente scientifica britannica di particolare vivacità, sebbene disuguale e discutibile nei suoi risultati, sorta a Cambridge tra gli anni sessanta e settanta intorno alla figura di Eric S. Higgs. Alcune questioni trattate nell'articolo non sono nuove, ma si stanno rivelando di insospettata importanza per ricostruire il comportamento dell'uomo all'epoca della sua definitiva emancipazione dai doni spontanei della natura: la fine del modo di vita paleolitico e l'inizio di quello neolitico. È in questo contesto che recenti ricerche italiane, ancora inedite o poco note, vengono a fornire un possibile contributo di fatti e di idee, non sempre nella direzione degli autori britannici.

Il messaggio fondamentale dell'articolo è che la «rivoluzione» neolitica non fu inevitabile. Che lo si scopra in Medio Oriente, o lo si percepisca nelle moderne revisioni della preistoria europea, non è altro che un dettaglio geografico: resta il fatto di un comportamento umano apparentemente generalizzato. A quanto pare il nuovo modo di vita basato sulla domesticazione della natura non fu una buona novella abbracciata con entusiasmo subito e dappertutto. Fu invece un mosaico di spunti vecchi e nuovi, accolto con indecisione e adattato con elasticità.

Ma dietro i fatti archeologici c'è probabilmente un messaggio ancora più generale, che molti autori britannici tendono a trascurare: caccia e allevamento non sono antitetici, ma si sovrappongono e oscillano. Nella storia del comportamento umano i rapporti tra l'uomo e gli altri animali sono stati assai più sfumati di quanto non si credesse fino a poco tempo fa. L'esitazione verso il modo di vita contadino, verso il Neolitico, dovrebbe ormai essere vista come un corollario del fenomeno più generale. Fino all'avvento della caccia come passatempo gratuito - in Egitto e in Assiria tremila anni fa, in Europa negli ultimi quattro o cinque secoli - caccia e allevamento non sono stati attività contrastanti: le comunità sono passate agilmente dall'una all'altra.

Laddove per oltre un secolo gli archeologi hanno visto una brusca opposizione c'è in realtà uno spettro continuo di variazione comportamentale, il cui fattore caratterizzante è ciò che chiamiamo l'«interesse alla pari» dell'uomo per l'animale. Ciò non toglie che l'affermarsi del modo di vita neolitico sia stato un processo di importanza decisiva nell'evoluzione recente dell'umanità. È anche chiaro che la mutua reversibilità di caccia e allevamento non va presa per identità: le due polarità economiche implicano ovviamente comportamenti, atteggiamenti e significati sociali diversi.

All'Istituto di antropologia dell'Università di Napoli, quelle che abbiamo chiamato le «oscillazioni caccia-allevamento» sono specifico oggetto di studio da quattro anni. Identificato il fenomeno come un problema antropologico, ciò che si cerca di fare è andare oltre il risultato immediato dell'analisi dei reperti faunistici e indagare il contesto, e quindi le possibili ragioni, delle oscillazioni stesse. Lo scrivente e alcuni collaboratori hanno riconosciuto e studiato il fenomeno in casi originali che spaziano dallo Yemen a una regione solo apparentemente più familiare, le nostre Alpi.

Nello Yemen, una recentissima scoperta italiana al margine del deserto ha rivelato episodi di caccia intensiva a equidi e gazzelle, che collimano con quanto osservato da Legge e Rowley-Conwy in Giordania e Siria. Si tratta delle prime informazioni del genere per l'intera Arabia centromeridionale. Interessanti differenze sembrano emergere rispetto al comportamento socioeconomico di alcuni gruppi dell'altopiano, legati all'allevamento bovino, all'interno di società «neolitiche» essenzialmente simili e contemporanee. Sull'altopiano alcuni indizi fanno presumere alternanze di caccia e allevamento fra gruppi e attraverso il tempo, ma le ricerche , svolte nell'ambito della missione italiana dell'IsMEO (Roma), sono solo all'inizio.

Nelle Alpi, le oscillazioni caccia-raccolta durante e dopo il Neolitico stanno risultando tutt'altro che banali come finora si è ritenuto, e lo stesso discorso comincia ad aprirsi per l'Appennino campano. Nei rapporti con gli animali, come in altri settori, l'acquisizione del Neolitico risulta avvenuta a mosaico, con disparità marcate tra valli maggiori e valli minori e fra gruppi delle stesse «culture». Inoltre, quello che Gordon Childe e alcuni studiosi svizzeri e tedeschi avevano già indicato come un «ritorno alla caccia», dopo due o tre millenni di presunto trionfo agricolo, verso la fine del Neolitico centroeuropeo, sembra avere compreso una quantità di oscillazioni di vario tipo, per le quali la regione alpina costituisce a nostro avviso un osservatorio privilegiato. Il caso più importante è in corso di studio a Breno, in Valcamonica, dove si è portata in luce la prima successione neolitico-calcolitica delle Alpi centrali (si veda il mio articolo *Il primo abitato neolitico delle Alpi centrali* in «Le Scienze», n. 205, settembre 1985), e un altro caso è stato forse riconosciuto in alta Val di Susa nelle Alpi piemontesi (Chiomonte).

La regione appenninica offre occasioni di studio non ancora sfruttate. Anche qui si può pervenire a interessanti prospettive teoriche, e comunque a una valutazione articolata e realistica delle relazioni tra l'uomo e i suoi animali. Su quest'area vertono le ricerche di Umberto Albarella, che sta per pubblicare un lavoro su un sito neolitico di cacciatori di cervi da poco scoperto al piede dell'Appennino avellinese.

Francesco Fedele

gli aspetti rimarchevoli del sito è il grado di dettaglio con cui tale processo può essere osservato, attribuibile in sostanza a due fattori. Uno è l'enorme ricchezza di reperti organici, sia animali sia vegetali, l'altro è la continuità di occupazione, ossia la presenza di una comunità agricola nello stesso luogo del precedente villaggio di cacciatori-raccoglitori.

Non è un caso che entrambi i villaggi si siano succeduti nello stesso luogo. L'ubicazione era certamente favorita da circostanze che incoraggiarono, fin da tempi remoti, un insediamento permanente su grande scala. Le piante della valle fluviale integravano la selvaggina della steppa nel fornire una disponibilità di cibo costante per tutto l'anno. Sul lungo periodo, la stabilità di questa economia di caccia e raccolta fu interrotta per due volte da intermezzi di rapido mutamento. Il primo caso fu l'introduzione della coltivazione dei vegetali; poi, un millennio dopo o poco più, in modo altrettanto brusco, il bestiame domestico sostituì gli animali che erano stati cacciati nella steppa.

Da questa vicenda, si possono trarre varie considerazioni. La prima è che l'avvento di un modo di vita basato sull'agricoltura non deve essere necessariamente un processo unitario. Un concetto molto dibattuto nell'archeologia degli ultimi decenni, quello di una «rivoluzione neolitica», vorrebbe associare il sorgere di insediamenti permanenti con l'origine dell'agricoltura; l'idea di una rivoluzione fa inoltre pensare a un cambiamento travolgente, venuto tutto d'un tratto. Ma la preistoria di Tell Abu Hureyra indica non tanto un mutamento unitario e radicale, quanto un processo per gradi, in cui gli elementi chiave - la vita sedentaria, la coltivazione, l'allevamento degli animali - furono adottati uno per volta in un arco di tempo piuttosto esteso.

C'è anche un altro aspetto che rende inadeguata l'ipotesi della «rivoluzione». Uno dei postulati strettamente connessi a tale ipotesi è che l'agricoltura sia intrinsecamente superiore alle economie di caccia e raccolta quanto a potere produttivo. Sul lungo periodo, questo è certamente vero. Ma in determinati ambienti, nella fase embrionale dell'agricoltura, la vecchia economia può risultare più vantaggiosa. Ad Abu Hureyra, la facile possibilità dell'uccisione in massa evidentemente ritardò l'avvento della domesticazione degli animali su vasta scala fino al momento in cui la produttività della caccia fu ridotta da influenze esterne. La domesticazione, in altri termini, non si affermò automaticamente né sostituì subito i vecchi metodi. Al contrario, essa si sviluppò in parallelo con quei metodi, finché non si dimostrò più adatta a rispondere a nuove necessità. Un grande contributo di Tell Abu Hureyra è l'avere aiutato gli archeologi a comprendere questo processo multiforme e di fondamentale rilevanza.

Inquinamento atmosferico da particelle in sospensione

Diversi metodi di analisi hanno permesso di dimostrare che il particellato atmosferico, che provoca riduzione di visibilità e danni a numerosi materiali, deriva in gran parte dall'impiego di combustibili fossili

di Robert W. Shaw

In gran parte del mondo le superfici esterne di edifici e sculture monumentali sono sottoposte a erosione; il terreno e le acque dei laghi hanno un'acidità superiore al normale, che provoca l'impoverimento o la scomparsa della vita vegetale e animale. Gran parte dei danni viene attribuita alle precipitazioni acide: pioggia o neve che contengono acidi in soluzione.

Tuttavia, le precipitazioni umide non costituiscono l'unico mezzo attraverso il quale gli inquinanti atmosferici raggiungono la terra; diffusione e altri processi permettono a gas e particelle acide di arrivare a terra in condizioni secche. Molti ambientalisti sospettano che i depositi secchi siano distruttivi, per i materiali e l'ambiente in genere, quanto le precipitazioni acide e che sarebbe opportuno ampliare la definizione di «pioggia acida» per includervi sia l'inquinamento umido sia quello secco.

I depositi secchi non sono frequente oggetto di discussione quanto le piogge acide, perché per molto tempo le conoscenze al riguardo sono state più scarse. Negli ultimi dieci anni, tuttavia, molti ricercatori hanno effettuato scoperte sorprendenti sulle fonti e i possibili effetti del particellato atmosferico. In effetti gli studi rivolti ad accertare l'origine delle particelle esaminate hanno permesso di confutare l'idea che emissioni naturali da aree paludose, da vulcani e dalla vegetazione fossero responsabili di gran parte delle ricadute acide in tutto il mondo. Attualmente, non vi è più dubbio che, anche in zone rurali, le deposizioni acide (sia umide sia secche) siano dovute quasi invariabilmente ad attività umane: principalmente all'impiego di combustibili fossili per produrre energia, per l'industria e per i trasporti.

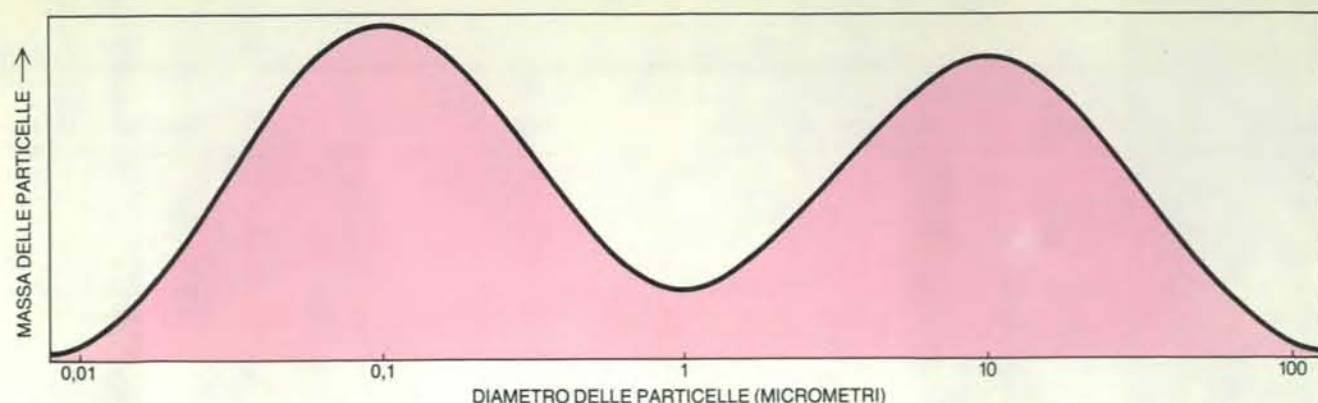
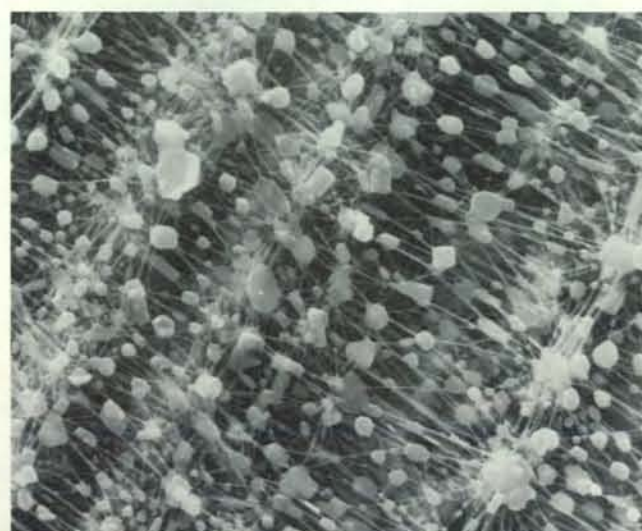
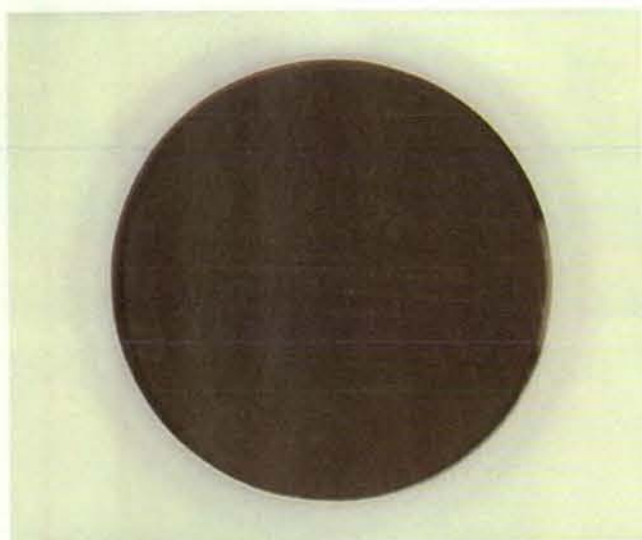
Il carbone e altri combustibili, bruciando, emettono varie sostanze, tra cui particelle di carbonio (se la combustione è inefficiente) e anidride solforosa (SO_2) gassosa. Le elevate temperature di combustione poi, fanno sì che l'azoto atmosferico si combini con l'ossigeno, per formare ossidi di azoto (NO_x), pure gassosi. Quando nell'atmosfera tali gas incontrano acqua o molecole affini formano goccioline di acido solforico (H_2SO_4) e acido nitrico (HNO_3) gassoso, entrambi facilmente solubili nell'acqua piovana. Se l'atmosfera è relativamente secca, l'acido nitrico tende a rimanere allo stato gassoso, mentre l'acido solforico forma minute particelle, che, a volte, possono raggiungere la terra con la pioggia, ma possono anche depositarsi in condizioni asciutte insieme ad altro particellato.

I primi studi sul particellato atmosferico non consentivano di fare piena luce sulla natura o le cause delle deposizioni acide, perché la misurazione e l'analisi delle particelle erano difficili. I campioni di acqua piovana contenuti nei recipienti di raccolta in genere riflettono la quantità e la composizione delle precipitazioni in una certa area, ma fino a che le particelle in sospensione non vengono raccolte con strumenti progettati allo scopo, i campioni risultanti non possono fornire informazioni attendibili. Inoltre, le particelle sono spesso una miscela di sostanze diverse, e diventa problematico identificarne i componenti e, di conseguenza, le fonti.

Nonostante questi ostacoli, la US Environmental Protection Agency (EPA), sul finire degli anni settanta, stabilì che il problema rappresentato dal particellato atmosferico era abbastanza importante da giustificare uno sforzo di ricerca nella raccolta e nell'analisi di campioni. Oltre a ricoprire di una patina scura gli edifici e a danneggiare i materiali esposti, si pensava che le particelle in sospensione contribuissero alle foschie che, durante l'estate, riducono la visibilità in vaste zone degli Stati Uniti orientali. Con l'aiuto di studiosi che si occupavano di progettazione di strumenti scientifici, di chimica analitica e di statistica, il nostro gruppo all'EPA iniziò a mettere a punto una serie di strumenti e di tecniche per raccogliere in modo efficace le particelle in sospensione, per determinarne la composizione e formulare ipotesi plausibili sui materiali che le producono.

Per effettuare un gran numero di prelievi in maniera riproducibile, avevamo bisogno di un campionatore automatico, che effettuasse la raccolta in modo tale da riflettere la concentrazione totale di particellato atmosferico in una certa area. Doveva anche dividere quanto raccolto in due frazioni, di particelle fini (di diametro inferiore a un micrometro) e di particelle grossolane, dato che studi precedenti avevano dimostrato che i due gruppi erano qualitativamente diversi.

Fu l'elaborazione statistica dei dati a indicare che la natura di particelle di dimensioni diverse poteva essere significativamente differente. Nei primi studi sulle deposizioni secche, i ricercatori effettuavano conteggi del numero di particelle in vari intervalli di dimensioni. Essi

L'atmosfera, fra le montagne della Georgia sovietica, era molto più limpida il 9 luglio 1979 (*in alto*) rispetto al 21 luglio (*in basso*). Le foschie e la visibilità ridotta nella seconda parte del mese furono prodotte da un cospicuo aumento della concentrazione di particelle contenenti solfato, le cui dimensioni fanno sì che esse diffondano efficacemente la luce. La presenza di solfato, raramente prodotto per processi naturali, fu sorprendente, perché dimostrava che l'atmosfera della regione era più inquinata di quanto si pensasse dalle emissioni di lontane aree industriali.

Una curva a due picchi rappresenta la distribuzione dei diametri delle particelle (in scala logaritmica) rispetto alla massa dei campioni. In ogni intervallo di dimensioni, l'area in colore sotto la curva è proporzionale alla massa dei campioni raccolti. L'andamento a due picchi, che compare invariabilmente nei campioni raccolti in ogni parte del mondo, riflette il fatto che le particelle di diametro inferiore a 1 micrometro differiscono per composizione e origine da quelle più grandi. Le tipiche particelle fini sono prodotti acidi della combustione, e si formano per accrezione, mentre quelle grossolane generalmente non sono acide e derivano dalla frantumazione meccanica del terreno e delle rocce.

Le frazioni fine e grossolana del particellato atmosferico hanno spesso un aspetto nettamente diverso. In un prelievo effettuato presso un cantiere di costruzione del North Carolina (*in alto*), la frazione fine (*a sinistra*), in gran parte prodotta dal gasolio delle macchine escavatrici, è nera, mentre la frazione grossolana (*a destra*), composta soprattutto da argilla frantumata dalle ruote dei veicoli, è gialla. Analogamente, in campioni prelevati dall'atmosfera cittadina (*in basso*) osservati al microscopio elettronico, le particelle fini (*a sinistra*) appaiono lisce e uniformi, ma quelle grossolane (*a destra*) sono frastagliate. La forma di una particella, vista a ingrandimento maggiore, può indicarne l'origine. Le microfotografie sono di David L. Johnson del College of Environmental Science and Forestry della State University of New York a Syracuse.

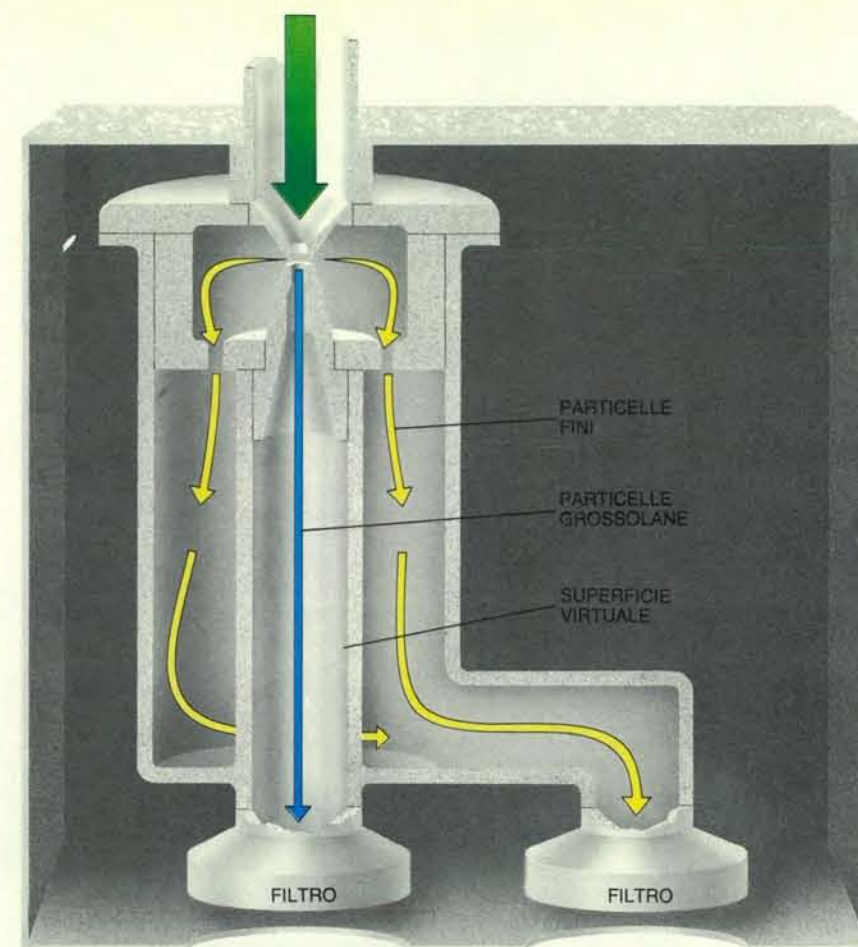
trovarono che il numero di particelle presenti aumentava al diminuire del loro diametro, un andamento certamente interessante, ma che non rivelava molto sulla natura dei campioni.

Poi, circa 20 anni fa, i ricercatori che stavano studiando lo smog a Los Angeles ebbero l'idea di considerare la massa totale del particellato nei diversi intervalli di dimensioni. In tutti i campioni, fra le centinaia in corso di studio, trovarono una divisione apparentemente naturale: le particelle con diametro compreso fra 0,1 e 1 micrometro costituivano la maggior parte della massa, mentre il resto era composto da particelle con diametro compreso fra 1 e 100 micrometri. Questo andamento costante suggeriva che le particelle fini e quelle grossolane non differissero solo per dimensioni, ma costituissero classi diverse di materiali.

Ulteriori studi hanno poi confermato questa conclusione. È stato dimostrato che nell'atmosfera le particelle fini derivano principalmente da processi chimici, in particolare reazioni di combustione, e sono essenzialmente acide. Perciò, esse richiedono l'esame più attento in molte analisi. Invece, le particelle grossolane sono prodotte soprattutto dalla disgregazione di materiali presenti in natura, come il terreno, e comprendono quasi esclusivamente sostanze non acide. I processi che regolano l'accrezione di una particella nell'atmosfera non sono ben compresi, ma sembra siano autolimitanti: l'accrescimento cessa quando il diametro si avvicina a 1 micrometro. I processi meccanici, d'altra parte, generalmente non possono produrre frammenti di diametro inferiore a un micrometro.

La difficoltà di raccogliere campioni rappresentativi e di dividerli in due gruppi è stata superata con la messa a punto di uno strumento chiamato conimetro a urto virtuale, che sfrutta la diversa tendenza di particelle fini e grossolane a cambiare direzione quando sono in movimento. Una corrente d'aria convoglia le particelle dall'ambiente allo strumento e qui si divide. Parte di essa continua a fluire in direzione rettilinea; la parte restante fa un angolo, muovendosi di 90 gradi lateralmente, prima di tornare a fluire nella direzione originale.

Le particelle più grosse comprese nella corrente sono troppo pesanti per deviare, perciò proseguono in linea retta con la corrente fino a una camera, detta superficie virtuale. Qui il loro movimento è rallentato, in modo da poter essere raccolte, senza urti, su un filtro di Teflon, distribuendosi uniformemente sulla sua superficie. (Nei primi strumenti, le particelle, in movimento veloce, spesso non si fermavano sui filtri, introducendo un errore nelle misurazioni.)

Le particelle più fini, invece, possono cambiare bruscamente direzione. Esse seguono le «curve» della corrente, e si raccolgono uniformemente su un secondo filtro, in un'altra zona del conimetro.

Il conimetro a urto virtuale è uno strumento di raccolta che cattura le particelle in sospensione in una corrente d'aria e le separa in due gruppi in base alle dimensioni. Le particelle più grosse, relativamente pesanti, non possono cambiare direzione quando parte della corrente devia lateralmente, ma continuano in linea retta fino a una camera (detta «superficie virtuale») dove il loro movimento è rallentato, in modo che non rimbalzino indietro dal filtro di raccolta. Gran parte delle particelle più piccole segue la corrente laterale e viene raccolta da un secondo filtro.

Funzionando automaticamente e potendo essere caricati con una serie di filtri di raccolta, tali strumenti possono raccogliere campioni per oltre un mese in assenza di operatori, secondo programmi prestabiliti. Abbiamo scelto filtri di Teflon perché la fibra di vetro, impiegata in origine, reagisce con i gas atmosferici in modo tale da creare nuove particelle e da confondere i risultati; il Teflon, invece, è inerte in queste condizioni.

Dopo che un campione è stato raccolto, se ne deve determinare la massa, per poter calcolare la concentrazione delle singole sostanze nell'atmosfera. La massa può essere misurata direttamente, pesando il filtro prima e dopo il prelievo del campione, ma tale procedimento è troppo lungo per essere esteso a molti campioni. Abbiamo perciò deciso di servirci di uno strumento in uso nella fisica nucleare, lo spessimetro a raggi beta. Il filtro con il campione viene posto fra una sorgente e un rivelatore di particelle beta. Lo strumento determina la massa contando il numero di elettroni che passa attraverso il filtro; quanto maggiore è la massa, tanto minore sarà il numero di elettroni che raggiungono il rivelatore.

Anche il secondo dei nostri test di routine, la spettroscopia di fluorescenza di raggi X, deriva dai metodi della fisica nucleare. In questo caso, raggi X, anziché particelle beta, vengono fatti passare attraverso il campione, per identificare gli elementi atomici. La radiazione stimola l'emissione da parte del campione di raggi X di vari livelli energetici, ognuno caratteristico di un particolare elemento e l'intensità della radiazione è una misura della concentrazione dell'elemento. Il metodo, uno dei molti per lo scopo, è assai sensibile e permette di misurare contemporaneamente la concentrazione di molti elementi; inoltre, come lo spessimetro a raggi beta, ha il vantaggio di essere non distruttivo.

La spettroscopia di fluorescenza X è una tecnica rapida e facile, e si presta a un impiego di routine, ma fornisce solo informazioni approssimative sulla composizione di un campione. Benché indi-

chi quali elementi compaiano, non può specificare, per esempio, se lo zolfo è presente come solfato (SO_4^{2-}), né se particolari elementi sono combinati nella stessa particella. Queste informazioni sono, però, importanti, perché lo stato chimico di un elemento determina il suo comportamento e il suo impatto sull'ambiente; inoltre, permette di ipotizzare la fonte dell'elemento in questione.

Per ottenere una maggiore specificità, sono necessarie altre analisi, alcune distruttive e altre no; la scelta fra le varie tecniche analitiche dipende dagli obiettivi del gruppo di ricerca e dai fondi a disposizione. L'esame al microscopio ottico ed elettronico è una valida metodologia non distruttiva, che richiede parecchio tempo, ma permette di identificare specie di particelle di morfologia nota.

Negli ultimi anni, alcuni ricercatori hanno talvolta impiegato l'analisi per diffrazione di raggi X dei campioni. La diffusione dei raggi X da parte di una sostanza cristallina rivela la sua struttura tridimensionale, che, a sua volta, può essere di guida nell'individuazione dell'origine del materiale in questione. Per esempio, questa tecnica ha permesso di evidenziare in alcuni campioni la presenza di mullite, un silicato di alluminio che si forma solo ad elevata temperatura, come nella combustione del carbone; la specifica struttura cristallina lo differenzia da altre forme di silicato di alluminio presenti nel terreno e nelle argille.

Oltre ai metodi distruttivi è generalmente necessaria anche l'analisi chimica di un campione per determinare lo stato dei suoi elementi costitutivi. Per esempio, la dissoluzione in acqua provoca dissociazione ionica, e permette di misurare la concentrazione di ioni idrogeno e quindi l'acidità totale del campione. (Secondo Arrhenius, un acido è una sostanza che libera ioni idrogeno in soluzione.) Questo tipo di analisi può dirci anche se l'azoto è presente come nitrato (NO_3^-) o come ammonio (NH_4^+), oppure se lo zolfo è presente come solfato.

Talvolta, è possibile dedurre la fonte di alcune particelle da queste manipolazioni, ma i ricercatori di solito devono basarsi su complesse analisi statistiche e su una buona dose di giudizio personale per giungere all'individuazione. Tecniche di elaborazione dei dati molto efficaci sono l'analisi fattoriale, sviluppata da psicologi nei primi decenni di questo secolo, e il «bilancio chimico degli elementi», introdotto all'inizio degli anni settanta nello studio delle particelle nell'atmosfera. L'analisi fattoriale contribuisce a identificare i materiali da cui gli elementi in un campione hanno origine, mentre il bilancio degli elementi indica il contributo relativo delle diverse fonti alla composizione del campione.

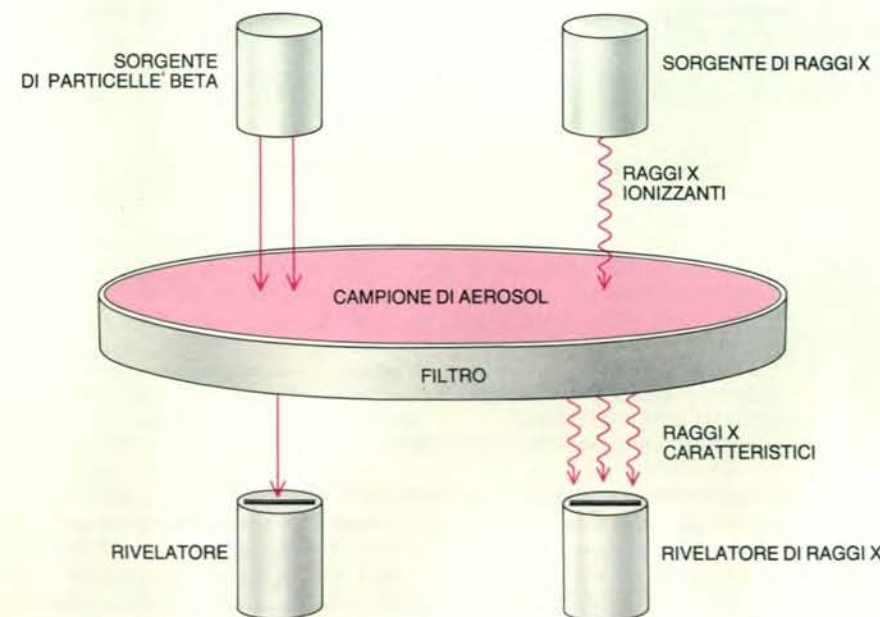
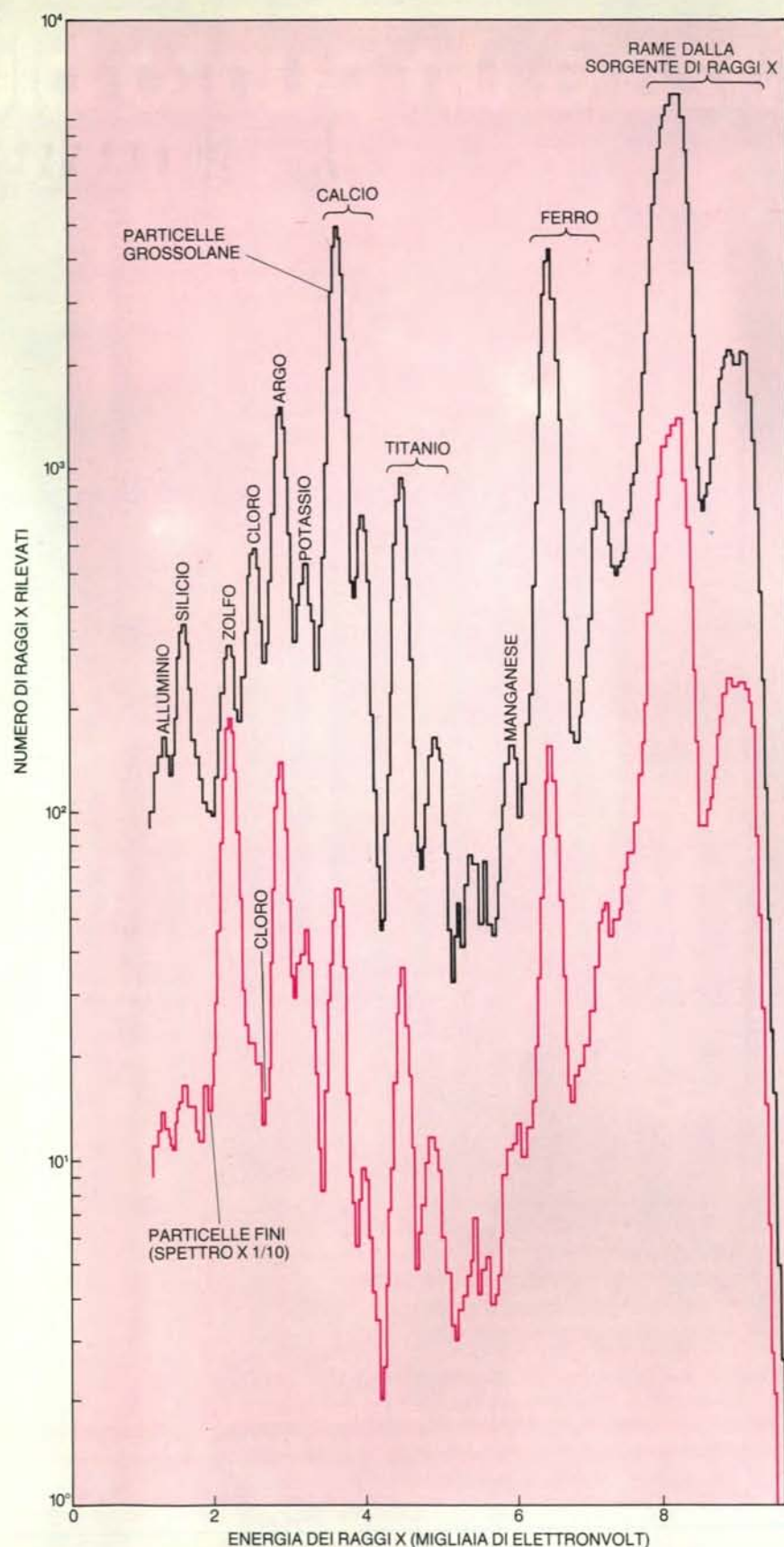
Nell'analisi fattoriale, il calcolatore passa in rassegna tutti i dati sulla composizione di campioni raccolti in momenti diversi, poi identifica gli elementi per i quali un aumento o una diminuzione di concentrazione appaiono correlati. Gli elementi le cui concentrazioni variano in modo analogo possono avere la stessa origine, e probabilmente sono presenti nelle stesse particelle. Per indicare quali materiali possono aver prodotto gli elementi correlati, il calcolatore confronta poi i gruppi di elementi ottenuti statisticamente con una lista di possibili fonti e con le «impronte digitali» di ognuna di esse, ossia con la tipica combinazione di elementi emessi.

Supponiamo che le concentrazioni di selenio e zolfo, misurate in quattro giorni consecutivi, siano pari al doppio del valore originale il secondo giorno, al triplo il terzo giorno e si riducano alla metà il quarto giorno. Il fatto che gli elementi varino allo stesso modo implica che possano provenire dallo stesso materiale. Un esame della lista delle fonti rivela poi che selenio e zolfo sono componenti del carbon fossile e suggerisce che questo materiale potrebbe essere responsabile della loro presenza nel campione.

Nel bilancio degli elementi, si compila una lista dei rapporti tipici fra gli elementi prodotti da un certo numero di fonti plausibili. Una lista di questo genere potrebbe indicare, per esempio, che il rapporto tra selenio e zolfo nel carbone di una regione è di 1 a 1000. Avendo misurato le concentrazioni degli elementi significativi in un certo campione, è possibile prevedere, in base ai rapporti standard nella lista delle fonti, le concentrazioni di altri elementi presenti, se tutti provenissero dalla stessa fonte.

Se in un campione sono stati ritrovati 0,01 microgrammi di selenio e la lista delle fonti indica che il selenio è un componente in traccia del carbone, si può supporre che tutto il selenio derivi dal carbone e che la quantità di zolfo della stessa origine dovrebbe essere di 10 microgrammi. Se, in effetti, nel campione è presente proprio questa quantità di zolfo, si può concludere che tutto il selenio e tutto lo zolfo provengono dal carbone.

Nella realtà lo stesso elemento può essere prodotto da fonti diverse. Il bilancio degli elementi fa fronte a questa complicazione esprimendo un certo numero di concentrazioni diverse di elementi, prodotte dalle probabili fonti. Si trova poi la «miscela» che meglio corrisponde alle concentrazioni misurate nei campioni. I calcoli potrebbero rivelare, per esempio, che l'80 per cento della frazione fine di un campione è un sottoprodotto della combustione del carbone e il 15 per cento è emesso dai veicoli a motore. Oppure, che il 94 per cento del piombo è prodotto dai veicoli a motore, il 4 per cento dall'incenerimento dei rifiuti e l'1 per cento dalla combustione del carbone.

Alcuni anni fa, dopo aver messo a punto strumenti e tecniche, il nostro gruppo dell'EPA impiantò un laboratorio temporaneo nel Great Smoky Mountains National Park, nel Tennessee. Il nostro compito, che si inseriva in uno studio di maggiore portata sulla qualità dell'aria nel parco, era di determinare il contributo relativo delle emissioni naturali della vegetazione e di quelle dei vei-

Due diverse tecniche analitiche permettono di analizzare le particelle senza rimuoverle dai filtri di raccolta. Per determinare la massa di un campione (*a sinistra*) si pone il filtro tra una sorgente di particelle beta e un rivelatore che ne effettua il conteggio. Al crescere della massa, il numero di particelle che riescono ad attraversare il campione diminuisce. Per identificare gli elementi presenti, si effettua quindi un'analisi separata mediante spettroscopia di fluorescenza di raggi X (*a destra*). I raggi X che attraversano il campione fanno sì che ognuno degli elementi emetta una fluorescenza X caratteristica. I livelli energetici dei raggi rivelano l'identità degli elementi presenti; l'intensità dei raggi (il numero dei raggi emessi) riflette le concentrazioni.

Lo spettrogramma X di particelle raccolte a St. Louis da Thomas G. Dzubay della US Environmental Protection Agency dimostra che lo zolfo è l'elemento più abbondante nella frazione fine. Lo zolfo nelle particelle fini deriva in genere dal carbone. (Il numero di raggi X emessi da ogni elemento della frazione fine è determinato moltiplicando per 10 il valore che compare in ordinata.) Il silicio, il calcio e il ferro, tipici componenti del terreno, costituiscono gran parte della frazione grossolana. Il picco dell'argo è prodotto dall'aria entro lo spettrometro.

coli a motore e dell'industria alla composizione del particellato atmosferico.

All'epoca, molti ritenevano che le foschie comuni in quella regione fossero causate da minuscole particelle, prodotte dalla reazione di idrocarburi gassosi emessi dalle piante sempreverdi con altre sostanze dell'ambiente. La fluorescenza X e studi di natura chimica dimostrarono, invece, che gli idrocarburi contribuivano poco alla massa dei nostri campioni, mentre il solfato acido risultava predominante. In effetti, i livelli di solfato erano molto maggiori di quelli tipici di ambienti allo stato naturale, inducendoci a concludere che le emissioni di qualche tipo di combustibile avevano influenzato l'atmosfera del parco.

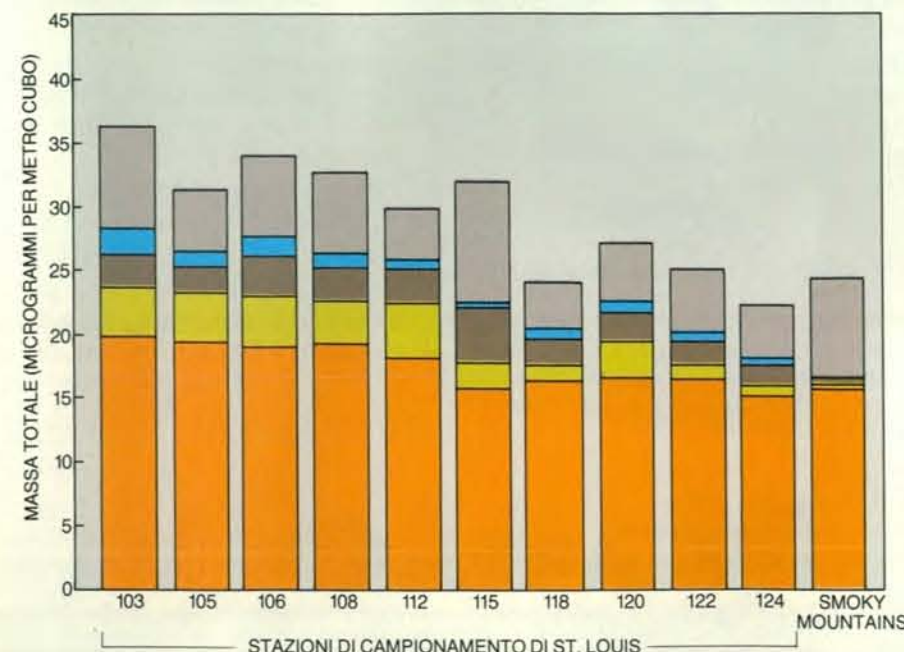
Sia i risultati da noi ottenuti nelle Smoky Mountains, sia uno studio analogo, nell'ambito di un programma congiunto di ricerche sull'ambiente tra Stati Uniti e Unione Sovietica, causarono molta sorpresa. L'obiettivo, in questo secondo caso, era di studiare la formazione di particellato atmosferico per opera dei composti chimici emessi dalle sempreverdi. Per ridurre il più possibile il disturbo dell'inquinamento atmosferico, scegliemmo come base un osservatorio astronomico isolato fra le montagne della Georgia sovietica, una regione famosa per la limpidezza dell'aria e per le splendide foreste di sempreverdi.

Per un mese, il nostro gruppo, insieme ad alcuni colleghi sovietici, misurò il tasso sia di gas, sia di particelle in sospensione. Lo studio dei gas rivelò la presenza di bassi livelli di anidride solforosa, ossidi di azoto e idrocarburi alogenati (la classe di composti a cui appartengono i refrigeranti responsabili della riduzione dello strato di ozono); tale risultato indica di solito una buona qualità dell'aria. La composizione della miscela di particelle in sospensione era, tuttavia, molto simile a quella riscontrata nelle Smoky Mountains: soprattutto solfato, con una quantità molto limitata di idrocarburi. In realtà, gli idrocarburi erano così scarsi che non riuscimmo a ottenerne abbastanza per compiere lo studio previsto.

Le quantità significative di particelle contenenti solfato in due zone ritenute immuni da inquinamento sollevano l'ovvio problema di quale sia la loro fonte principale. Noi riteniamo che possa trattarsi della combustione di carbón fossile in regioni lontane, perché le emissioni, in questo caso, vengono generalmente rilasciate nell'atmosfera da alte ciminiere. Questo espediente permette di ridurre l'inquinamento nelle zone immediatamente circostanti, ma fa sì che l'anidride solforosa gassosa rimanga nell'atmosfera abbastanza a lungo da allontanarsi dal luogo di origine e da combinarsi con altre sostanze per formare particelle. Conformemente a questa teoria, abbiamo trovato che le particelle di solfato nella Georgia sovietica, una zona posta a grande distanza da concentrazioni industriali, sono più «vecchie» di quelle

nelle Smoky Mountains. Negli Stati Uniti orientali, circa un quinto dello zolfo presente nell'atmosfera è contenuto in particelle; nella Georgia sovietica, vi è 30 volte più zolfo sotto forma di particelle che anidride solforosa gassosa.

Benché sospettassimo che il carbone fosse la fonte principale del solfato presente nelle Smoky Mountains e presso la nostra base in Unione Sovietica, potemmo dimostrare definitivamente il collegamento tra centrali termiche e inquinamento atmosferico da particelle solo nel corso di uno studio di 16 mesi condotto nella valle del fiume Ohio: qui vi è la più alta concentrazione di centrali termiche a carbone di tutti gli Stati Uniti.

Nel nostro studio, volevamo esaminare aree che non fossero sicuramente compromesse; stabilimmo perciò le stazioni di rilevamento nelle campagne del Kentucky, dell'Indiana e dell'Ohio, lontano da città, grandi strade e dalle emissioni di fumo delle centrali termiche.

La foschia sulle Great Smoky Mountains (*in alto*), visibile nella fotografia di Robert K. Stevens dell'EPA, è dovuta soprattutto a particelle di solfato nell'aria. Un tipo di analisi statistica, nota come bilancio degli elementi chimici, che dà indicazioni sulle fonti delle particelle in base alla composizione dei campioni, mostra che gran parte del solfato nella regione deriva dalla combustione del carbone in centrali termiche lontane. L'analisi indica anche (*in basso*) che la concentrazione di particelle derivate dal carbone nelle Smoky Mountains non è minore di quella rilevata in 10 stazioni di raccolta a St. Louis, città altamente industrializzata. I dati sono stati raccolti da Thomas G. Dzubay.

Trovammo che, in media, circa il 50 per cento della massa di particelle fini era costituito da solfato; la concentrazione totale era quasi pari a quella riscontrata nelle città industriali della regione. Potemmo stabilire così un collegamento certo con la combustione del carbone, per l'associazione costante tra la concentrazione del solfato e quella di un elemento in traccia, il selenio. Infatti, non ci sono altre probabili fonti di selenio nella valle del fiume Ohio o nei dintorni.

Oggi è ormai certo che il solfato costituisce la maggior parte della massa di particelle fini presenti nell'atmosfera degli Stati Uniti orientali e di altre regioni. Poiché le particelle contenenti solfato sono spesso estremamente acide, è molto probabile che anch'esse, come le piogge acide, siano in grado di danneggiare materiali e di alterare l'equilibrio acido-base degli ecosistemi.

Vi sono inoltre indizi che queste particelle, per le loro dimensioni e le conseguenti proprietà di diffusione della luce, possano influenzare l'equilibrio termico tra l'atmosfera e la superficie terrestre, provocando perturbazioni meteorologiche. Per una strana coincidenza, le particelle fini nell'atmosfera hanno dimensioni tali da diffondere la luce con elevata efficienza. Calcoli basati su modelli meteorologici indicano che questa diffusione può provocare il raffreddamento della superficie terrestre e, nello stesso tempo, riscaldare l'alta atmosfera. Queste alterazioni possono, a loro volta, modificare i moti naturali delle masse d'aria e stimolare la formazione di inversioni termiche, ossia di andamenti in cui la temperatura dell'aria aumenta, invece di diminuire, all'aumentare della quota. Le inversioni termiche provocano il ristagno dell'aria e degli inquinanti che essa contiene.

Benché tali effetti debbano ancora essere dimostrati con certezza, un altro evidentissimo fenomeno è ormai ben noto: sempre a causa della diffusione della luce, alte concentrazioni di particelle ricche di solfato nell'atmosfera riducono la visibilità. La conferma che il solfato contribuisce al fenomeno giunge da diversi studi, fra cui quelli compiuti dal nostro gruppo nella Georgia sovietica e nella valle del fiume Ohio. Un'analisi di livelli di visibilità e consumo di carbone negli Stati Uniti rivela che la visibilità diminuisce in maniera significativa con l'aumento del consumo di carbone. (Negli ultimi decenni, vi è stata riduzione di visibilità in molte regioni del globo, persino nelle zone artiche.)

Nella Georgia sovietica, trovammo che la visibilità diminuiva chiaramente all'aumentare della concentrazione di particelle di solfato. Nella valle del fiume Ohio, la percentuale di zolfo presente in forma di particelle è di gran lunga maggiore in estate, parallelamente a un aumento nell'incidenza di foschie. Probabilmente, la concentrazione di solfato aumenta d'estate perché vi sono periodi più lunghi di luce e temperature e umidità superiori. Queste condizioni accelerano la conversione dell'anidride solforosa gassosa in solfato. Nello stesso tempo, gli episodi più frequenti di stagnazione dell'aria permettono agli inquinanti di accumularsi nell'atmosfera.

Il solfato è il componente dominante del particellato fine nell'atmosfera di gran parte degli Stati Uniti, ma un recente studio, della durata di un anno, compiuto nella zona di Los Angeles dimostra che anche il carbonio può essere un componente importante della frazione fine; in effetti, a Los Angeles, rappresenta il costituente principale della massa di tali particelle. Il carbonio in questa forma sembra essere un sottoprodotto diretto dei combustibili liquidi che alimentano veicoli a motore e industria e non derivare dagli idrocarburi emessi dagli alberi o da altre fonti. Le particelle di carbonio possono essere emesse, inoltre, non dalle moderne centrali termiche a combustibile liquido, ma da processi di combustione meno efficienti, come quelli che si verificano nelle stufe a legna diffuse in molte case di campagna.

Le particelle di carbonio sono una miscela di nerofumo (carbonio puro) e molecole organiche complesse. Oltre a diffondere la luce e a diminuire la visibilità analogamente alle particelle di solfato, esse assorbono la luce, facendo sì che le foschie contenenti tali particelle appaiano scure. Il nerofumo in sé non sembra dannoso per la salute, ma alcune molecole organiche che vi aderiscono possono esserlo. D'altra parte, le molecole organiche potrebbero svolgere almeno un ruolo potenzialmente positivo, in quanto esse sono probabilmente caratteristiche dei vari tipi di combustibili che le producono e possono perciò indicare la fonte del carbonio.

Gli studi compiuti dal nostro e da altri gruppi hanno dimostrato che solfato e composti del carbonio prodotti per combustione costituiscono, in genere, i componenti principali della frazione fine del particellato atmosferico in qualsiasi grande regione geografica. Talvolta, però, si rileva la presenza di sostanze inattese, che i metodi descritti sopra permettono di spiegare adeguatamente.

Alcuni anni fa, in dicembre, scoprimmo con sorpresa che il cloro rappresentava circa il 20 per cento della massa di particelle fini nei campioni raccolti in varie località del Middle West per un periodo di alcuni giorni. Il cloro, che si trova nell'aria di regioni costiere sotto forma di sale marino, normalmente contribuisce alla composizione delle particelle in zone continentali per non più dello 0,1 per cento. Nondimeno, le analisi effettuate nel Wisconsin, Kansas, Missouri, Kentucky, Indiana e Ohio indicarono tutte un aumento di cloro vistoso e praticamente simultaneo.

Incuriositi, passammo in rassegna diverse fonti potenziali di cloro. Il sale (cloruro di sodio) che viene sparso sulle strade dopo una nevicata era una possibile spiegazione e i livelli di cloro nei nostri campioni erano compatibili con tale origine. Tuttavia, non vi erano state nevicate, né spargimento di sale presso le località di campionamento. Anche la combustione di sostanze vegetali può produrre un aumento del cloro, che dovrebbe però essere accompagnato da un analogo incremento, da noi non rilevato, nei livelli di potassio e di nerofumo. Le misure eseguite su altri composti indicavano che, per molti aspetti, l'aria nei siti di raccolta era insolitamente pulita: la concentrazione di ozono e di ossidi di zolfo e di azoto era molto più bassa del normale, e così pure la concentrazione di solfato nella massa di particelle fini.

Questi dati suggerivano una conclusione sorprendente: anziché rappresentare uno strano episodio di inquinamento, il cloro indicava un'intrusione, ancora più insolita, di aria pulita. I calcoli eseguiti al National Center for Atmospheric Research confermavano i risultati, suggerendo che una massa d'aria marina proveniente dall'oceano Pacifico settentrionale avesse attraversato il Canada e fosse penetrata nel Middle West.

Un altro fatto analogo mise alla prova l'abilità deduttiva di Thomas G. Dzubay dell'EPA. Dopo aver raccolto molti campioni di particellato atmosferico a St. Louis, egli notò che alluminio, potassio, stronzio, antimonio e bario erano molto più abbondanti in uno dei campioni rispetto a tutti gli altri. Ricordando che tutti questi elementi sono presenti nei fuochi d'artificio, Dzubay controllò le date dei campionamenti: come previsto, il campione «strano» era stato raccolto tra mezzogiorno e mezzanotte del 4 luglio (festa nazionale negli Stati Uniti).

Questo episodio dimostra che la complessità delle particelle, un tempo grave impedimento per le analisi, è in realtà un vantaggio: osservando quali elementi o composti sono presenti contemporaneamente, si può spesso dedurre da quali materiali abbiano avuto origine. Forse, fra qualche anno, le tecniche di analisi saranno abbastanza precise da accertare non solo che l'uso di carbone o di altri combustibili è responsabile della presenza di particellato atmosferico in una certa regione, ma anche dove è avvenuta la combustione.

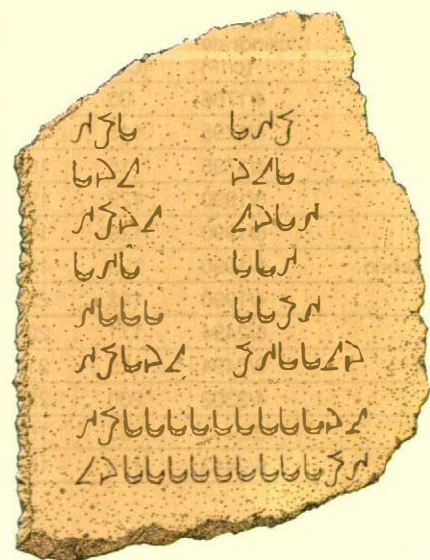
Senza dubbio, comunque, l'enorme quantità di dati raccolti permette di affermare, fin da ora, che l'uso di combustibili è la fonte principale di precipitazioni acide e di deposizione di particelle, e che le masse d'aria inquinata superano i confini nazionali. Ora è necessario che i responsabili politici, i dirigenti industriali e i cittadini, in tutto il mondo, affrontino il problema di determinare cosa bisognerà fare per migliorare i controlli già in atto sull'inquinamento, quanto denaro dovrà essere speso a questo scopo e da chi.

(RI)CREAZIONI AL CALCOLATORE

di A. K. Dewdney

Scale di parole e una torre di Babele portano a inattingibili vette di calcolo

Nessuno sa come o perché sia scomparsa la civiltà marziana, ma una caverna ai piedi del Mons Olympus conserva un misero resto di questa antica e meravigliosa cultura. In fondo alla caverna, avvolta in una stoffa polverosa di ignota composizione, c'è una tavoletta che ricorda la famosa stele di Rosetta. Tolta la stoffa si rivela quello che sembra essere una specie di vocabolario: un elenco di parole sulla metà sinistra della tavoletta è messo in corrispondenza con un elenco di parole sulla destra. Nella parte bassa della pietra sono incise due frasi collegate tra loro in modo particolare. La prima può essere trasformata nella seconda attraverso ripetute sostituzioni di parole tratte dal vocabolario: ogni parola che si trova nella colonna di sinistra del vocabolario può essere sostituita dalla corrispondente parola nella colonna di destra.

Martiansentencescontainnospaces (gli enunciati marziani non contengono spazi). In effetti, non ce n'è bisogno dato il carattere fluido ed espressivo dell'Antico Marziano: ci sono molti modi per dividere una frase marziana in parole. Come regola generale, però, solo poche divisioni hanno senso in un dato contesto. Si può illustrare il processo di sostituzione utilizzando la prima frase di questo paragrafo. Se nella colonna di sinistra del vocabolario comparisse *tainnospaces* e *cealwisdom* fosse la corrispondente parola sulla destra, si potrebbe sostituire questa parola alla prima per ottenere una nuova frase valida: *Martiansentencesconcealwisdom* (gli enunciati marziani contengono saggezza).

I saggi dell'antico Marte ritenevano che si potessero ottenere tutte le nozioni degne di essere apprese partendo da una frase fondamentale e sostituendo le parole secondo il vocabolario della saggezza, di cui sopravvive solo il frammento che si trova nella caverna. La frase fondamentale è la prima delle due incise in basso sulla tavoletta del Mons Olympus raffigurata in questa pagina. In generale, un calcolatore non sarà in grado di verificare se si può derivare una certa frase da quella fondamentale. In altri termini non si può scrivere un programma per calcolatore (per quanto grande o veloce esso sia) che decida correttamente, per ciascun vocabolario e ciascun insieme di due parole (o frasi) in ingresso, se è possibile tradurre la prima parola nella seconda. Spiegherò oltre il perché. Nel frattempo, sono capaci i lettori di trasformare la frase fondamentale in quella che appare sotto di essa? (Questa particolare trasformazione può essere compiuta.)

Il problema della traduzione del marziano fa parte di un'intera famiglia di rompicapo che hanno come scopo la trasformazione di parole, frasi e anche interi paragrafi in altre parole, frasi e paragrafi. I primi sviluppi dell'argomento si devono, in parte, al reverendo Charles Dodgson, altrimenti noto come Lewis Carroll. Tra i molti passatempi matematici e simbolici inventati da Carroll durante le sue passeggiate per Oxford c'è una trasformazione chiamata «scala di parole». In una scala di parole si parte con due parole della lingua che si è scelta. La prima parola viene detta «sorgente» e la seconda «obiettivo». Si può trasformare la parola sorgente nella parola obiettivo, cambiando una lettera per volta?

Dizionario marziano di saggezza

Se le due parole hanno lo stesso numero di lettere, il compito è banale. Ma si può avere la certezza che tutte le successioni di caratteri intermedie siano a loro volta parole? Si tratta di tutt'altra questione, ma è pur sempre una difficoltà che gli accademici del gioco di parole dovrebbero poter agilmente superare. Per illustrare il procedimento, si può, in effetti, partire da *horse* (cavallo). In un passo la parola si può trasformare in *house* (casa) e un altro passo porta a *mouse* (topo). Con questa alchimia sembrerebbe di poter trasformare un cavallo (*horse*) in quasi tutto ciò che ha un nome di cinque lettere. È possibile?

Per dare una risposta alla questione in termini generali è necessario costruire una rete di trasformazioni. Per non usare una terminologia eccessivamente d'effetto, l'ho ribattezzata «ragnatela di parole». Costruire la ragnatela completa per la lingua inglese sarebbe una cosa mostruosa: ogni parola vi verrebbe rappresentata come un singolo punto. Per soddisfare i requisiti della scala di parole due punti sono congiunti da una linea se le parole corrispondenti differiscono esattamente di una lettera. Per esempio, per trovare le parole nelle quali si può trasformare *horse*, si traccia un percorso attraverso la ragnatela di parole a partire da *horse* inserendo in una lista tutte le parole incontrate lungo il cammino. Un'esplorazione completa di questi percorsi sarebbe analoga ad afferrare la ragnatela per il punto *horse* e contemporaneamente tirare tutti i punti a esso collegati. L'insieme avrebbe probabilmente l'aspetto di una rete da pesca malamente rammendata. Naturalmente la grande maggioranza delle parole non verrebbe «tirata su» insieme a *horse*; tutte le parole che non sono formate da cinque lettere, per esempio, rimarrebbero al di sotto.

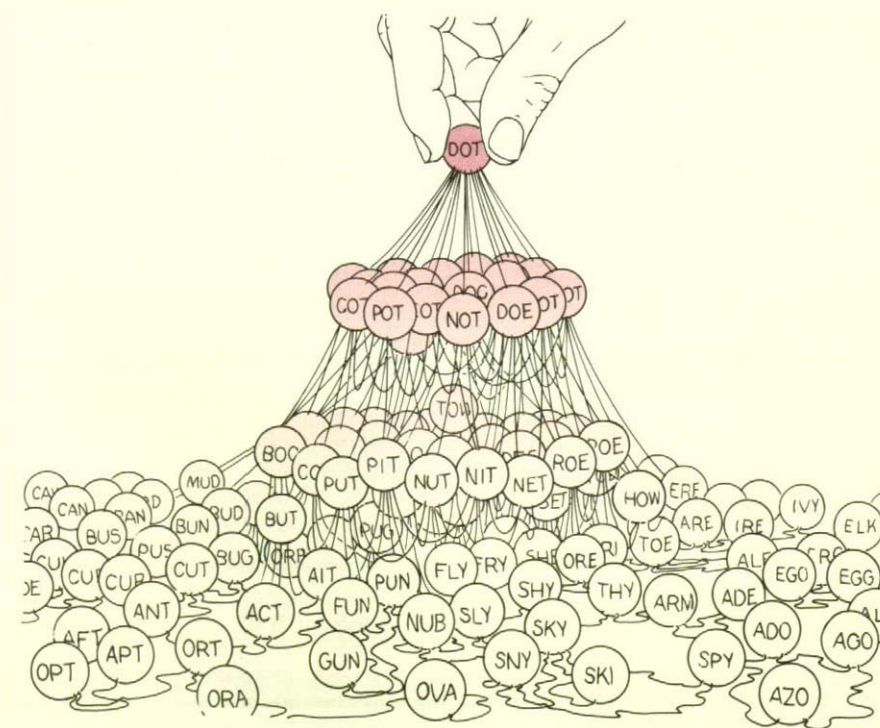
Verrebbero fuori tutte le parole di cinque lettere? Vale a dire, è possibile costruire una scala di parole che parta da *horse* e comprenda tutte le altre parole inglesi di cinque lettere? Non so come rispondere a questa domanda, ma, a giudicare dalle risposte avute in passato da parte dei lettori, sarei pronto a scommettere che qualcuno la troverà. È necessario avere un dizionario presentato in forma leggibile dalla macchina e un programma che cerchi nel repertorio le parole connesse. Ogni volta che una parola viene aggiunta alla lista di tutte le parole collegate a *horse* il programma la cancella dal dizionario in memoria, fa passare di nuovo l'intera lista confrontando ciascuna parola in essa con ciascuna delle parole del dizionario in costante diminuzione. Ogni volta che una parola del dizionario differisce da una parola della lista per una lettera, viene aggiunta alla lista e il procedimento ricomincia. Dizionari diversi porteranno sicuramente a risultati diversi. O no? Forse. D'altra parte, due dizionari ragionevolmente completi potrebbero dare lo stesso risultato. Detto in altre parole, le ragnatele possono essere così fitte che la loro topologia risulta praticamente identica.

Alla domanda se parole di n lettere siano o meno connesse, si può rispondere anche senza l'aiuto della macchina per piccoli valori di n. Sicuramente, all'interno di una ragnatela, tutte le parole di una lettera sono collegate. Anche il caso di parole di due lettere si può decidere immediatamente. Secondo il mio dizionario Scrabble le parole di due lettere sono *aa, ad, ae, ah, ai, am, an, ar, as, at, aw, ax, ay, ba, be, bi, bo, by, da, de, do, ef, eh, el, em, en, er, es, et, ex, fa, go, ha, he, hi, ho, id, if, in, is, it, jo, ka, la, li, lo, ma, me, mi, mu, my, na, no, nu, od, oe, of, oh, om, on, op, or, os, ow, ox, oy, pa, pe, pi, re, si, so, ta, ti, to, um, up, us, ut, we, wo, xi, xu, ya* e *ye*.

Basta poco per accorgersi che tutte le parole della ragnatela sono collegate. Prima di tutto, tutte le parole di due lettere che iniziano con la stessa lettera sono reciprocamente collegate. Questo ci permette di raggruppare insieme tutte le parole che iniziano con la stessa lettera. Rimane solo da osservare che il gruppo delle *a* è collegato con gli altri quattro gruppi di vocali, che il gruppo delle *b* è collegato con ciascuno degli altri 16 gruppi consonantici e che il gruppo delle *a* e quello delle *b* sono tra loro collegati.

Sulla base di queste due sole prove relative a casi di parole con un basso numero di lettere, alcuni potrebbero spavaldamente ipotizzare che tutte le parole inglesi di n lettere siano collegate. Tale affermazione è un tipico esempio di quell'esaltazione che dà il gioco di parole astratto. I lettori dotati di un calcolatore possono cercare di verificare l'affermazione per un altro piccolo valore di n come $n = 3$ (*si veda l'illustrazione di questa pagina*). Quelli che non hanno il calcolatore devono accontentarsi di trasformare *hate* (odio) in *love* (amore) o, ancora meglio, di passare da *evil* (male) a *good* (bene), compiendo notevoli passi avanti nella comprensione morale e logo-logica del procedimento.

Il gioco delle scale di parole costituisce un semplice esempio di un cosiddetto sistema di trasformazione testuale. I due ingredienti principali di un simile sistema sono un insieme di testi da trasformare e una regola di trasformazione. I testi potrebbero essere parole, frasi o paragrafi, che sono in realtà semplici successioni di caratteri, indipendentemente da come viene scelto l'insieme di testi. La regola di trasformazione consiste in una procedura per sostituire una parte di una successione data con un'altra successione. Nel seguire la procedura in genere si consulta un dizionario.

Nelle scale di parole, per esempio, l'insieme di successioni consiste di tutte le parole inglesi di un dato dizionario. La procedura prende ogni parola a cui si applica e ne genera un'altra che differisce dalla prima per una lettera e che si trova sempre nel dizionario. In questo caso il dizionario è un semplice elenco di parole.

La ragnatela di parole di tre lettere

La maggior parte dei dizionari pubblicati hanno voci che consistono di parole «definite» in termini di altre parole. Qui è implicito un altro tipo di catena di trasformazione, più complesso della scala di parole. Si immagini, per esempio, di partire con una parola inglese, cercarne il significato e fare un elenco di tutte le parole che si trovano nella sua definizione. Poi si guardino i significati di quelle parole, si faccia un elenco composto delle parole in essi contenute e si continui a ripetere il procedimento. Dato che l'inglese ha un numero finito di parole (cosa che non potrei garantire per le altre lingue), il procedimento sviluppa subito una caratteristica, per così dire, ciclica: certe parole cominciano a ricorrere. Esse sono, in parte, definite sulla base di se stesse. Le si potrebbe chiamare parole primarie. Non vanno subito sbandierate pronunciandole a gran voce, ma vanno ripetute in silenzio con una sorta di timore: forse costituiscono l'indefinibile fondamento concettuale dell'inglese.

Il progetto, o qualcosa di molto simile, è stato svolto da Robert Amsler del Bell Communications Research. Amsler si è servito di un dizionario modificato in modo da ridurre la definizione di ogni parola a un insieme di parole, indicanti il «genere», che servono a identificare le ampie classi di cose alle quali la parola appartiene. Alle parole di genere sono accostate differenze: parole che servono a distinguere una parola data da altri elementi del genere a cui appartiene. Utilizzando un programma analogo a quello descritto prima, Amsler ha percorso il suo dizionario speciale e ha trovato fondamentalmente quello che i filosofi del linguaggio avevano previsto che avrebbe rinvenuto: parole primitive usate tutte nelle definizioni finali di altre parole primitive. Alcuni esempi: *food* (cibo), *person* (persona), *thing* (cosa), *instrument* (strumento) e *group* (gruppo). Le parole primitive occupavano i rami finali di quella che è stata chiamata una gerarchia intricata, qualcosa di simile a una foresta in cui i rami di alberi diversi siano cresciuti insieme. Le parole primitive si trovano «in su», in direzione di una sempre maggiore generalità.

L'idea fondamentale di Amsler è che i dizionari sono più strutturati e rivelano di più sul linguaggio di quanto comunemente si pensi. Sarebbe sorprendente se lo schema relativamente semplice «per genere e differenza» non portasse ad alcune anomalie. Un *barbecue*, per esempio, risultava «un animale arrostito su un fuoco all'aperto». Il *barbecue* si veniva così a trovare in un elenco in qualche punto tra *aardvark* (oritteropo) e *zebra*.

La ricerca di Amsler trova applicazione negli schemi di recupero delle informazioni da basi di dati, nei quali l'utente vuole recuperare casi particolari di una classe generale. Attualmente Amsler sta ricostruendo la sua gerarchia intricata sul *McGraw-Hill Dictionary of Scientific and Technical Terms*.

Una catena di trasformazioni, che ha rapporti con quella di Amsler, è stata sperimentata da Ron Hardin, ricercatore presso gli AT & T Bell Laboratories di Murray Hill, New Jersey. Hardin, nel

afraid – apprehensive – expectant – hopeful – confident – **unafraid**
classified – secret – mysterious – unidentified – **unclassified**
dress – groom – tidy – clear – wipe – take off – divest – **undress**
even – level – uniform – regular – periodic – spasmodic – **uneven**
perfect – pure – unvarnished – unfinished – rough – **imperfect**
potent – dynamic – animated – excited – nervous – weak – **impotent**
rational – clever – witty – funny – silly – brainless – **irrational**
safe – undamaged – mint – untarnished – polished – slippery – **unsafe**
social – friendly – peaceable – inoffensive – retiring – **antisocial**
valid – convincing – plausible – specious – unsound – **invalid**

La torre di Babele di Ron Hardin

corso degli anni, ha ideato parecchi giochi di trasformazione testuale interessanti e divertenti. La sua ultima preoccupazione riguarda il *New Collins Thesaurus*. Generalmente si serve di un repertorio lessicale chi cerca sinonimi di una parola o di una classe concettuale di parole che ha in mente. Si potrebbe immaginare che ogni parola si trovi su una nuvoletta di parole che significano tutte praticamente la stessa cosa e che le nuvole non siano tutte nettamente separate l'una dall'altra, ma che alcune si sovrappongano. Ne deriva che, benché certe coppie di parole non siano esse stesse reciprocamente sinonime, possono avere sinonimi in comune. E qui sta la fonte di molti problemi. Si considerino le seguenti catene scoperte da Hardin. Ciascuna parola si trova nella nuvola di sinonimi della parola che la precede:

acceptable → *so-so* → *ordinary* → *inferior* → *rotten* → *unacceptable* [accettabile, così così, ordinario, inferiore, sgradevole, inaccettabile]

reliable → *steadfast* → *obstinate* → *wayward* → *unpredictable* → *unreliable* [affidabile, risoluto, ostinato, ribelle, imprevedibile, inaffidabile]

Hardin ha generato parecchie migliaia di catene di questo tipo con il calcolatore. In ogni catena ciascuna parola viene trasformata, per passi successivi, nel suo contrario. Per lo più le catene generate da Hardin non sono più lunghe di questi esempi. Questo esperimento ci ricorda che i linguaggi umani sono fluidi, che il significato delle parole dipende pesantemente dal contesto e che ogni parola è carica di una certa ambiguità che è, a un tempo, una benedizione e una maledizione. Si consideri il caso di un avvocato che sottoponga un testimone a un contro interrogatorio riguardo al carattere di un imputato:

«Nella sua deposizione lei ha affermato che Watson è un uomo affidabile. Non lo definirebbe anche «risoluto?»

«Sì.»

«Non direbbe forse che, di tanto in tanto, l'imputato era ostinato nella sua risolutezza?»

«Beh, penso di sì.»

«Certe volte, la sua ostinazione si risolveva sicuramente in una forma di ribellione, non è vero?»

«Non ne sono sicuro.»

«Non ne è sicuro? Andiamo, dottor Finch, essere ostinato e ribelle è praticamente la stessa cosa.»

«Suppongo di sì.»

«Un ribelle è un uomo imprevedibile; in breve un uomo inaffidabile. Lei ha appena contraddetto la sua deposizione iniziale.»

Potete vedere, in alto in questa pagina, un ulteriore campionario di successioni per passare da una parola al suo contrario. L'uso che Hardin fa del *New Collins Thesaurus* non è così semplice come io l'ho presentato. Il repertorio lessicale è ordinato per parole principali. Quando l'utente vuole una nuova parola per un dato contesto cerca la parola principale con il significato più vicino a quello desiderato. Sotto la parola principale sono elencate le parole a essa collegate ordinate in base al significato. Per esempio, sotto la parola principale *express* si possono trovare significati che hanno a che fare col discorso e con la velocità. Il problema principale di Hardin era quello di collegare i significati in maniera logica e coerente attraverso le parole principali. Per dirla con Hardin: «Se si suppone che le distinzioni che vengono fatte quando i significati sono distinti non debbano scomparire quando vengono estese... non c'è alcun modo coerente per collegare il repertorio.»

Hardin, perciò, ha ideato un algoritmo che esamina sistematicamente tutte le coppie di parole principali tra loro collegate e calcola la forza del collegamento tra i vari significati in cui ciascuna parola può essere usata. I significati con il grado di collegamento più alto vengono quindi inseriti in una sorta di ragnatela di parole intessuta dall'algoritmo. In questo modo le distinzioni tra i significati tendono a rimanere ben definite, anche se qualche volta diventano confuse. Quando la ragnatela è completa, viene impiegato un secondo algoritmo per trovare il percorso più breve tra due significati qualsiasi. A quel punto è stato facile per Hardin identificare diverse coppie di parole per creare la propria torre di Babele, dal nome del biblico edificio che non poté essere portato a termine a causa della confusione delle lingue. La torre di Hardin è un mini-volume che presenta speciali attrattive per linguisti e avvocati.

In questa rubrica, nel novembre 1985, era stata citata una precedente trasformazione testuale di Hardin. In quell'occasione citai una poesia intitolata «Topeka Beagle Buffers», una trasformazione testuale della filastrocca per bambini «Peter Pipper». Il segreto della sua costruzione non è concettualmente più difficile di quello che sta alla base della torre di Babele.

Ecco come Hardin ha trasformato il verso che comincia con «'Twas the night before Christmas» (era la notte prima di Natale), per esempio, in un altro che comincia con «Tweeze denied beef worker isthmus». Prima ha scomposto tutte le parole del verso originale in una successione di fonemi: suoni base raggruppati secondo grossolane somiglianze. Avendo eseguito la stessa operazione per un dizionario standard, ha usato un calcolatore per raggruppare i fonemi in nuove parole, ignorando i confini tra le parole del testo originario quando necessario. Prima che la follia fosse completa, Hardin e alcuni colleghi dei Bell Laboratories hanno dovuto risistemare alcuni punti troppo rozzi della versione al calcolatore. La poesia che ne risulta lascia chi si appresta a celebrare il Natale con un'idea stranamente confusa di quello di cui parlava l'originale.

Non resta ora che ritornare al problema della traduzione marziana. Dato un dizionario di sostituzioni possibili e due parole (lunghe a piacere), perché è impossibile a un calcolatore stabilire in termini generali se si può ricavare la seconda parola dalla prima attraverso una successione di sostituzioni ricavate dal dizionario? Il problema viene detto «problema della parola di Thue», dal nome del matematico norvegese Axel Thue.

Molte persone al di fuori del settore dei calcolatori (e anche alcune addette) non si rendono conto che ci sono cose che i calcolatori semplicemente non possono fare. Una delle prime persone ad asserire che i calcolatori hanno limitazioni intrinseche fu il padre fondatore della scienza dei calcolatori, l'inglese Alan M. Turing. Turing ideò un semplice tipo di calcolatore astratto, chiamato (non da lui) macchina di Turing. Il congegno è teoricamente capace di eseguire qualunque calcolo che un calcolatore, attualmente esistente o allo stato di progetto, sia in grado di compiere. Turing dimostrò che per una macchina di Turing è impossibile decidere, data un'altra macchina di Turing e il suo ingresso, se questa seconda macchina si fermerà o no. In termini più concreti, è impossibile scrivere un programma che, dato un secondo programma e dei dati come ingresso, decida se il secondo programma smetterà mai di manipolare quei dati.

Il teorema di Turing ebbe almeno un'applicazione pratica. Ci fu in passato

LE RISORSE MINERARIE

LE SCIENZE *edizione italiana di* **SCIENTIFIC AMERICAN**

ha dedicato all'argomento diversi articoli:

L'origine dei depositi metalliferi nella litosfera oceanica
di E. Bonatti (n. 116)

La produzione mondiale di petrolio
di A.R. Flower (n. 117)

La produzione mondiale di carbone
di E.D. Griffith e A.W. Clarke (n. 127)

Le risorse mondiali di uranio
di K.S. Deffeyes e I.D. MacGregor (n. 139)

L'ipotesi dell'esistenza di gas nelle profondità della Terra
di T. Gold e S. Soter (n. 144)

Coltivazione di petrolio
d. R.A. Dick e S.P. Wimpfen (n. 148)

Per una strategia razionale dell'esplorazione petrolifera
di H. W. Menard (n. 151)

L'oro invisibile
di G. Facca (n. 165)

La microbiologia nell'estrazione mineraria
di C.L. Brierley (n. 170)

Giacimenti minerari da sorgenti termali oceaniche
di P.A. Rona (n. 211)

L'industria europea di minerali e metalli
di F. Ippolito e B. Curli (n. 226)

un esperto di sistemi di un certo istituto scolastico che decise che un grande calcolatore veniva utilizzato troppo a lungo per eseguire programmi di studenti dei primi anni. Uno dei motivi principali sembra consistesse nel fatto che molti programmi contenevano cicli infiniti in cui i principianti cadevano per errore. Se si fosse potuto esplorare il programma, prima dell'esecuzione, con un programma di ricerca dell'infinito, si sarebbe risparmiato molto tempo. Purtroppo, come un allievo di Turing si premurò di informare l'esperto, un programma del genere era impossibile.

Considerata l'irresolubilità del cosiddetto problema della fermata, è relativamente semplice dimostrare all'uomo della strada che anche il problema della traduzione marziana è irrisolvibile. Si procede dimostrando che il problema della fermata si può trasformare in quello della traduzione marziana.

Operare la trasformazione è concettualmente semplice. Al calcolatore viene fornita una lunga successione di simboli che descrive essenzialmente, in una opportuna codifica, i contenuti iniziali della memoria del calcolatore e dei suoi registri di lavoro. Viene poi costruita una seconda successione che indica che il calcolatore è in uno stato di fermata: ha completato tutti i calcoli che doveva svolgere. Il programma che deve essere sottoposto a verifica per stabilire se si fermerà viene tradotto in una serie di regole per trasformare la prima successione nella seconda per mezzo di sostituzioni. Ciascuna successione intermedia di simboli rappresenta, allora, un passo nell'attività del calcolatore mentre procede, si spera, a completare l'analisi dei dati che gli vengono presentati. Le regole si servono della sostituzione per far passare questo calcolatore virtuale da stato a stato seguendo il programma.

Se, per esempio, si deve immagazzinare in una certa locazione di memoria il contenuto di un registro, le regole sostituiscono il nuovo contenuto di quella locazione di memoria nella parte della successione di simboli che rappresenta la locazione di memoria in questione. Il procedimento di trasformazione riuscirà a ottenere la seconda successione dalla prima se e solo se il calcolatore in questione si ferma, completando l'analisi dei dati forniti. Poiché, secondo il teorema di Turing, il calcolatore non si fermerà mai, la trasformazione non si può fare.

Anche senza un'argomentazione così fantasiosa, l'irresolubilità della traduzione marziana viene suggerita dall'osservazione seguente: se la parola sorgente è trasformabile nella parola obiettivo, un programma che tenti sistematicamente ogni sostituzione concepibile arriverà alla fine a quelle giuste. Ma se la trasformazione è impossibile, lo si scoprirà mai? I lettori che non riuscissero a risolvere la nostra speciale esemplificazione del problema della traduzione marziana dovranno aspettare un mese.

In luglio ho parlato dei modi in cui i calcolatori vengono impiegati in borsa. La discussione si articolava in tre stadi, a partire dalla periferia e procedendo verso il cuore del mercato. Per prima cosa, i singoli operatori che hanno accesso al calcolatore possono acquistare uno qualsiasi delle centinaia di programmi dedicati all'analisi tecnica o fondamentale delle azioni. Più vicino alla borsa, i grandi istituti di investimento e gli agenti di cambio si servono dei calcolatori per pianificare l'acquisto o la vendita di azioni «future» (l'opzione di acquistare o vendere un'azione in una data futura e a un prezzo stabilito nel presente). Più vicino ancora, c'è la possibilità di sostituire lo *specialist* di borsa, totalmente o in parte, con un programma per calcolatore. Uno *specialist* è chi controlla tutta la contrattazione di una certa azione in una certa borsa. Il ruolo dello *specialist* può essere assunto da un programma per calcolatore che controlli tutte le transazioni d'acquisto e di vendita, cambiando via via il prezzo in modo adeguato? Ho fatto appello agli «esperti» tra i nostri lettori e non sono rimasto deluso.

Aaron C. Brown è un consulente finanziario di New York che qualche anno fa ha scritto una tesi di laurea sull'analisi tecnica nel settore finanziario, all'Università di Chicago. Mette in guardia dall'imbarcarsi nell'analisi tecnica, contro la quale gli sono sembrati un po' troppo deboli i miei ammonimenti: «Ci sono prove schiaccianti che [queste analisi] non funzionano, nel senso di produrre dei profitti; peggio ancora, sono pericolose.» Particolarmente pericolosa, dice Brown, è la strategia di comprare un'azione quando il prezzo scende in previsione di una risalita a breve termine. Questa strategia, avverte, è in stretta parentela con il sistema di raddoppio di puntata a ogni perdita in un ciclo di scommesse. Un investitore potrebbe così comprare per tutto il lungo declino di un'azione, sperando ogni volta in un'impennata del prezzo.

Charles J. Higgins, del Department of Finance and Computer Information Systems della Loyola Marymount University di Los Angeles, ha esaminato un procedimento tecnico particolare chiamato «tecnica dei diagrammi della linea di tendenza», simile alla tecnica del canale di contrattazione di cui ho parlato in luglio. A giudizio di Higgins, ciò che fa la differenza sono i costi di transazione. Nei dati sui prezzi studiati da Higgins, la maggioranza dei titoli non produceva guadagni anomali significativi rispetto alla media del mercato. «Tuttavia, molti titoli producevano guadagni significativamente anomali... quando i costi di transazione erano bassi o nulli.» Questo, dice Higgins, significa l'opportunità per gli *specialist*, gli operatori alle contrattazioni e altri membri della borsa (che pagano poche tasse di transazione o addirittura non le pagano affatto) di ottenere significativi guadagni.